# Mathematical Challenges in Electron Microscopy

**Robert James Tovey**

Supervisors: Dr. M. Benning
Prof. C.-B. Schönlieb

Department of Applied Mathematics and Theoretical Physics
University of Cambridge

This dissertation is submitted for the degree of
*Doctor of Philosophy*

Fitzwilliam College

September 2020

# Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee

Robert James Tovey
September 2020

# Mathematical Challenges in Electron Microscopy

Robert James Tovey

Development of electron microscopes first started nearly 100 years ago and they are now a mature imaging modality with many applications and vast potential for the future. The principal feature of electron microscopes is their resolution; they can be up to 1000 times more powerful than a visible light microscope and resolve even the smallest atoms. Furthermore, electron microscopes are also sensitive to many material properties due to the very rich interactions between electrons and other matter. Because of these capabilities, electron microscopy is used in applications as diverse as drug discovery, computer chip manufacture, and the development of solar cells.

In parallel to this, the mathematical field of inverse problems has also evolved dramatically. Many new methods have been introduced to improve the recovery of unknown structures from indirect data, typically an ill-posed problem. In particular, sparsity promoting functionals such as the total variation and its extensions have been shown to be very powerful for recovering accurate physical quantities from very little and/or poor quality data. While sparsity-promoting reconstruction methods are powerful, they can also be slow, especially in a big-data setting. This trade-off forms an eternal cycle as new numerical tools are found and more powerful models are developed.

The work presented in this thesis aims to marry the tools of inverse problems with the problems of electron microscopy: bringing state-of-the-art image processing techniques to bear on challenges specific to electron microscopy, developing new optimisation methods for these problems, and modelling new inverse problems to extend the capabilities of existing microscopes. One focus is the application of a directional total variation to overcome the limited angle problem in electron tomography, another is the proposal of a new inverse problem for the reconstruction of 3D strain tensor fields from electron microscopy diffraction data. The remaining contributions target numerical aspects of inverse problems, from new algorithms for non-convex problems to convex optimisation with adaptive meshes.

# Acknowledgements

This journey started in January 2014 with a nervous email sent to a young Dr Schönlieb who was the lead academic of a small group in my local university. I wanted to experience the working life of a 'proper mathematician' during my summer holiday. I met with Jan Lellmann who reassured me that, if I liked mathematics and computers, then image processing was a great place to be. I can say with certainty that without Carola and Jan, and especially their welcome and encouragement to me, I would not be here today. Their actions are by no means the default and I cannot thank them enough for their selfless generosity.

In 2015 I returned to Cambridge for my master's and became reacquainted with Carola and Martin Benning, whom I had also met in 2014. There is not much to say about this year, indeed a 'year' of study in Cambridge is only 16 weeks long, so I will discuss those people who became my PhD supervisors. Carola's support, loyalty, and friendship has been un-faltering since we first met. She always puts her all into every one of her responsibilities; a near infinite list. She is a wonderful person whose friendship I value greatly. Martin has also been a great friend these last four years, the phrase 'life of the party' does not quite do him justice. Moving away from Cambridge, I will miss him greatly and I look forward to catching up again soon. At least these new 'Neural Networks' have stopped him saying the word 'Bregman' so much…

There are too many people to thank individually. Discussion with other 'Cambridge maths' people has always been very stimulating. I have learned a lot in these last four years; mathematically, culturally, and about the world in general. It is the people of Cambridge that I will miss the most and, to all of you who have shared my journey, I thank you for the experience you have given me and wish you all the very best for your own futures. I will add two visitors to this list. Ozan Öktem, thank you for your expertise, patience, and insight. Stephan Hilb, working with you taught me how much fun collaboration can be. Discussing mathematics, computing, and politics with you was a complete joy. I wish you every success, wherever you end up in the world.

Finally, I would like to thank Paul Midgley and his group in the materials science department. Everyone has welcomed me and been prepared to give me endless time to teach me physics. I have learned a huge amount and my collaboration with you will always stand out as a highlight of my PhD.

My journey towards this point began nearly seven years ago. In that time I have finished three degrees, got married, possibly emigrated, experienced brexit, and I am now living through a global pandemic. I look forward to discovering a new journey.

# Table of contents

# Chapter 1

# Introduction

## 1.1 Overview

At the core of this thesis is the theme of combining the mathematics of inverse problems with the application of electron microscopy (EM). The purpose of this first chapter is to provide a general background to allow for a non-expert audience. Section 1.2 gives a conceptual grounding in both fields to motivate what are the core problems and why they are of practical importance. This leads to Sections 1.3 and 1.4 which expand upon the mathematical and physical details respectively. Once the principle concepts have been introduced, Section 1.5 will outline the main contributions of this thesis and the structure of the remaining chapters.

## 1.2 General background

### 1.2.1 Inverse problems

If causality is the principle of ‘cause and effect’, then *inverse problems* is the study of *reconstructing* a cause from a measured effect, the corresponding *forward problem* is to compute the effect of a given cause. Such an outlook can categorise almost any task, from the everyday to the extraordinary. Whether it is using the shadow of a tree to compute the time of day or detecting gravitational waves, humans have been solving inverse problems for millennia. On the other hand, deciding whether a problem is the forward or inverse can sometimes be a point of philosophy. When one looks at a digital clock, is one reading the time directly (forward problem), or matching the light arriving at your eyes to a set of memorised digits (inverse problem)?

A more mathematical distinction between forward and inverse problems might highlight the concepts of *well-/ill-posed* problems and data corruption. The modern statement of Hadamard’s well-posedness conditions can be found in Engl et al. (1996). An inverse problem is called well-posed if for all admissible data:

1. solutions exist,
2. solutions are unique,
3. and the solution depends continuously on the data.

These terms are highly abstract, in particular the term ‘solution’, but the philosophy agrees with the consensus that well-posedness depends on the uniqueness and stability of inversion. If only the continuity condition is violated then Engl et al. (1996) also suggest an intermediate term *‘mildly–ill-posed’* depending on more specific properties of the inverse problem.

If a problem is ill-posed then the literature of inverse problems provides powerful methods for extracting desired information from given data. Algorithms are designed to be stable against errors in the data and provide guarantees that the unique solution should be close to the true solution.

The canonical analytical expression for an inverse problem is:

$$\text{find the best} \quad u \quad \text{such that} \quad \mathrm{F}(u) \approx \eta \tag{1.1}$$

where $\mathrm{F}\colon \mathbb{U} \to \mathbb{V}$ denotes the forward problem and $\eta$ is the observed data. It is assumed that there is some *ground truth* $u^\dagger$ which generated $\eta$ through the model F, i.e. $\eta \approx \mathrm{F}(u^\dagger)$. The inverse problem is to find a reconstruction $u^*$ which satisfies $\mathrm{F}(u^*) \approx \eta$ in the appropriate sense and is a ‘good’ approximation of $u^\dagger$.

In the modern era, perhaps the most characteristic and intuitive class of inverse problems is that of photography. In this case, the standard analytical formulation is:

- $u^\dagger \in L^1(\mathbb{R}^2)$ is a 2D scene,
- $\eta \in \mathbb{R}^{m\times m}$ is a photograph,
- and F models the properties of the photodetectors in the camera.

Each of us are exposed to photos every day; many of which will be edited or processed in some way before we see them. Common image manipulations include the core areas of *denoising*, *super-resolution*, and *inpainting*. These will appear later in this thesis and represent characteristic examples of well-posed, mildly ill-posed, and ill-posed inverse problems respectively. Concrete examples are given in Sections Image denoising, Image inpainting, and Image super-resolution, however, to interpret these in the setting of electron microscopy, we also need the concept of indirect measurement.

Whether an inverse problem is direct or indirect is dictated by F. A photo of a 2D scene is direct, because each pixel in $\eta$ is like a point evaluation of $u^\dagger$, whereas if each pixel of $\eta$ were to represent a Fourier coefficient of $u^\dagger$, then it would be an indirect problem. There is no precise definition of each case but if the link between $u^\dagger$ and $\eta$ can be seen by the naked eye (of a non-specialist) then the problem is direct. Reading blurred text is typically considered a direct problem but automated subtitles (a map between sounds and strings of characters) is indirect. In a more scientific environment, devices including radar, medical CT, ultrasound and MRI machines all record indirect measurements of the object of interest.

Thankfully, computers struggle much less than humans to interpret indirect measurements, and the complexity of any given task can still be interpreted with the same concepts of denoising, super-resolution, and inpainting. The key new concept is that of *reconstruction artifacts*. For instance, medical CT and MRI are both modalities for mapping the body's internal organs, however, each observes different types of indirect measurements. As a result of this, when errors are made in reconstructing the organs they follow characteristic patterns which are easily recognised by specialists. In both examples, F is a linear map therefore reconstruction artifacts are elements of the null-space of F. Recognising reconstruction artifacts in each application is equivalent to learning the characteristic structures of each null-space. From the perspective of inverse problems, the extra challenge introduced by indirect measurements is to design reconstruction methods which are also robust to these intrinsic reconstruction artifacts.

The main type of indirect measurement seen in this thesis is electron tomography. The technical definition and etymology of the word tomography suggest it covers any technique capable of visualising 'slices' of structure hidden under the surface. A more common-tongue definition is that tomography is the process of recovering full information from many averaged observations. As an example, one (2D) photo of a (3D) man from the side does not tell you how broad he is but with multiple 2D photos from different angles we can see the full 3D scene. This is still not quite tomography because cameras only see the outer surface of an object. If the camera is replaced with an X-ray machine then we can observe all of the skeleton and internal organs, however, any single 2D image still does not have depth perception. Combining multiple X-ray images into a 3D model of an object is the simplest form of tomography. Note that 'X-ray transform' is just the name of a mathematical model which will be defined formally in Section 1.4.6, it does not imply that the physical modality uses X-rays. Indeed, the X-ray transform is the most common model for electron tomography and some 2D examples of this inverse problem are given in the Section Indirect measurements.

### Image denoising

Noise is observed whenever there are modelling inaccuracies. Without noise, the analytical problem is to find $u^*$ such that

$$\mathrm{F}(u^*) = \eta = \mathrm{F}(u^\dagger).$$

The most common noise model is stochastic additive noise where we assume

$$\eta = \mathrm{F}(u^\dagger) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}$$

for some known noise distribution $\mathcal{N}$. In standard imaging, such noise arises from the finite counting of photons (Poisson noise) and imperfect sensors (Gaussian white noise). Other noise models can be considered, for instance Rudin et al. (2003) consider multiplicative noise which accounts for pixels with different (unknown) levels of sensitivity.

Figure 1.1 shows the results of some classical algorithms for image denoising. Fourier-based methods have been used for over 50 years, for example by Schwartz and Shaw (1975). These are motivated by the idea that natural images are quite smooth (low frequency components) whereas stochastic noise is very oscillatory (high frequency). In this example we choose a value of $r > 0$ such that:

$$\mathcal{F}[u^*](\boldsymbol{k}) = \begin{cases} \mathcal{F}[\eta](\boldsymbol{k}) & |\boldsymbol{k}| < r \\ 0 & \text{else} \end{cases}, \qquad \text{and} \qquad \|u^* - \eta\|_2 = \left\|u^\dagger - \eta\right\|_2.$$

Removing noise by dampening the high frequency Fourier components is effective but typically results in over-smoothed images.

Variational methods such as Total Variational (TV) denoising have been under active development for over 30 years (Rudin et al., 1992; Benning and Burger, 2018) and are able to remove noise without loosing some sharp features in the image. In this example the TV reconstruction is defined as

$$u^* \in \operatorname*{argmin}_u \left\{ \|\nabla u\|_1 \quad \text{s.t.} \quad \|u - \eta\|_2 \leq \left\|u^\dagger - \eta\right\|_2 \right\}.$$

When $u$ is finite dimensional, $\nabla u$ can be approximated in many ways, c.f. Condat (2017); Chambolle and Pock (2020). Lebrun et al. (2012) suggest that the most common denoising problems can be considered well-solved since a wave of very good patch-based methods were developed. Dabov et al. (2007) proposed the BM3D algorithm which is still a very successful example of patch-based methods.

### Image inpainting

Inpainting is the task of filling in details of images which are missing from the original data. A simple example is to solve $\mathrm{F}(u^*) = \eta$ where

$$\eta(\boldsymbol{x}) = u^\dagger(\boldsymbol{x}) \qquad \text{for all } \boldsymbol{x} \in \Omega.$$

The set $\mathbb{R}^2 \setminus \Omega$ is called the inpainting domain.

Inpainting is an ill-posed inverse problem because we wish to see details in $u^*$ which are not visible in $\eta$. For example, if we have a photo where someone has their eyes shut, then we may wish to render the corresponding image with their eyes open. In this case the inpainting domain is the union of the two eyes. The exact eye colour is always unknown, therefore we

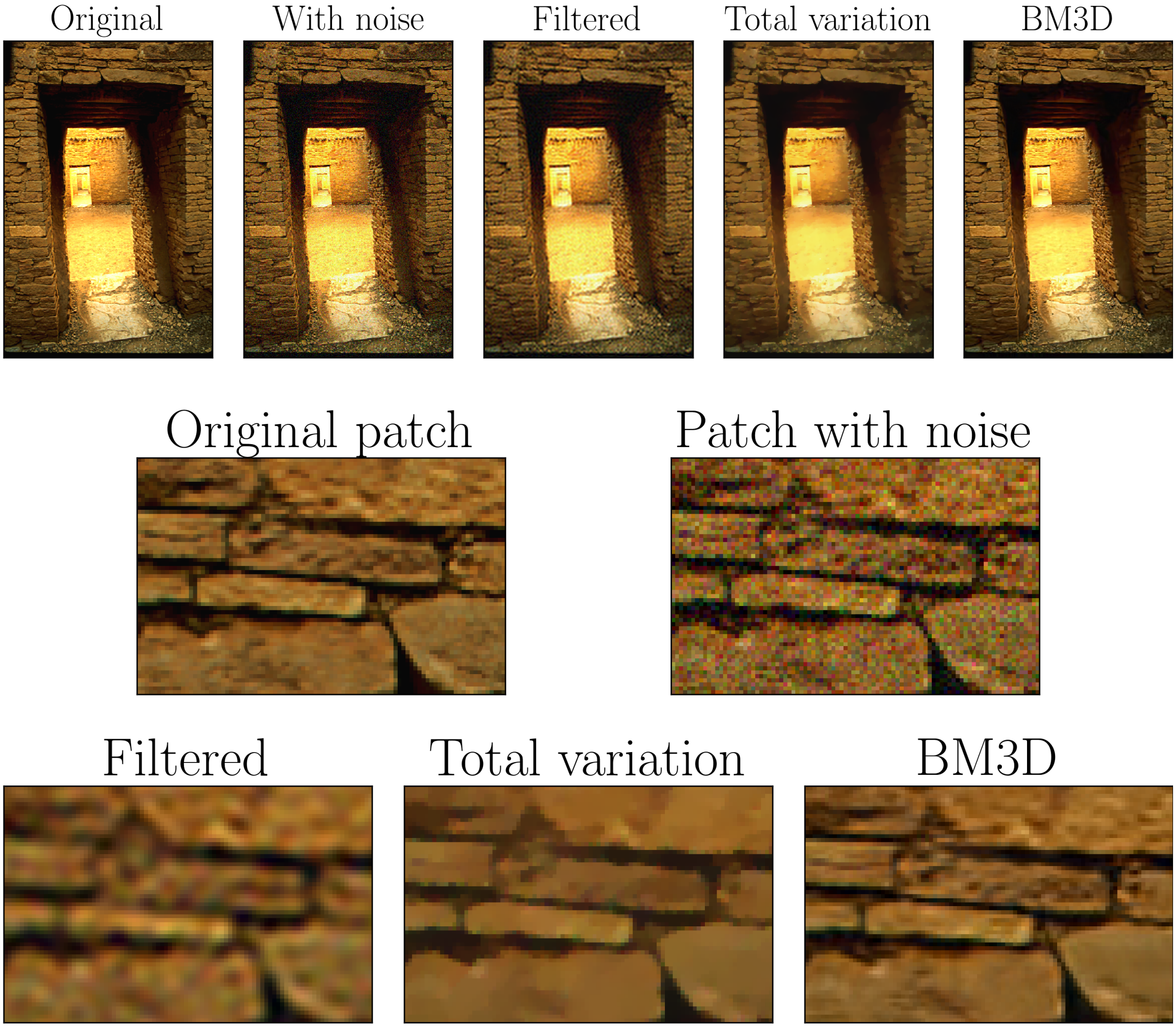


**Figure 1.1** Image from Robert E. Barber[a]. 7% Gaussian white-noise is added to an original, high-quality image. The filtered reconstruction removes all high-frequency Fourier coefficients from the noisy image. Total variation is a variational approach which encourages smoothness in the final image[b]. BM3D is a patch-based approach encourages self-similarity in local neighbourhoods[c].

[a] Barber Nature Photography REBarber@msn.com

[b] http://www.ipol.im/pub/art/2013/61/

[c] http://www.ipol.im/pub/art/2012/l-bm3d/

cannot guarantee to match $u^\dagger$ exactly without further information. This makes it very difficult to satisfy the uniqueness condition for well-posedness.

One characteristic of an inpainting problem is that it is not possible to expect a correct answer, but that there are certainly many wrong answers to avoid. For instance, eyes should not be bright pink nor a generic flesh coloured smear. Humans are typically very quick to notice incongruous visual details, therefore inpainting has remained a task where it is very challenging to achieve performance equivalent to human intuition.

Figure 1.2 shows an example with a rectangular inpainting domain within an image of pills. Each reconstruction looks fine upon a casual inspection, however, closer inspection reveals problems with each. In both cases, the rectangular outline of the inpainting domain is still visible. The transition is too sharp in the non-local reconstruction and too blurred in the learned reconstruction, neither of these properties were in the original image. We also notice when pills visible in the full image are not faithfully reproduced inside the inpainting domain. There are:

- two red and one white disc which enter the domain but do not complete the circle inside,
- and two red/white pills which start red outside of the domain and never appear inside the domain.

Again, we emphasise that there is not a correct answer, but that wrong answers can be quite noticeable. More details of the exact methods can be found following the links in the caption of Figure 1.2 but are not relevant to the remainder of this thesis.

### Image super-resolution

Super-resolution is the task of creating (realistic) fine structures which interpolate observed coarse structures. The canonical example is the desire to zoom in to an image beyond the resolution of the raw photo. In this example, it is similar to a highly structured form of inpainting where $\Omega = \{\boldsymbol{x}_{i,j}\}$ is a coarse mesh and the data is $\eta_{i,j} = \fint_{\omega_{i,j}} u^\dagger(\boldsymbol{x})d\boldsymbol{x}$ where $\omega_{i,j} = \{\boldsymbol{x} \in \mathbb{R}^2 \text{ s.t. } |\boldsymbol{x} - \boldsymbol{x}_{i,j}|_\infty = \min_{\boldsymbol{y}\in\Omega} |\boldsymbol{x} - \boldsymbol{y}|_\infty\}$.

Figure 1.3 demonstrates the performance of two very different methods. Bicubic splines have been used for image interpolation since the work of Hou and Andrews (1978), however splines themselves date back to the work of Schoenberg (1946). Splines perform a linear and continuous interpolation on $\mathbb{R}^2$ given by the formula

$$u^*(\boldsymbol{x}) = \sum_{i,j=1}^{m} \eta_{i,j} k(\boldsymbol{x} - \boldsymbol{x}_{i,j})$$

where $k \colon \mathbb{R}^2 \to \mathbb{R}$ is the *spline function*. There is a natural trade-off between smoothness and locality. For example, for Bicubic interpolation $k$ is piecewise $C^3$ and is supported on region covering $4 \times 4$ pixels, whereas a Bilinear spline is piecewise $C^1$ and supported on a $2 \times 2$ square.

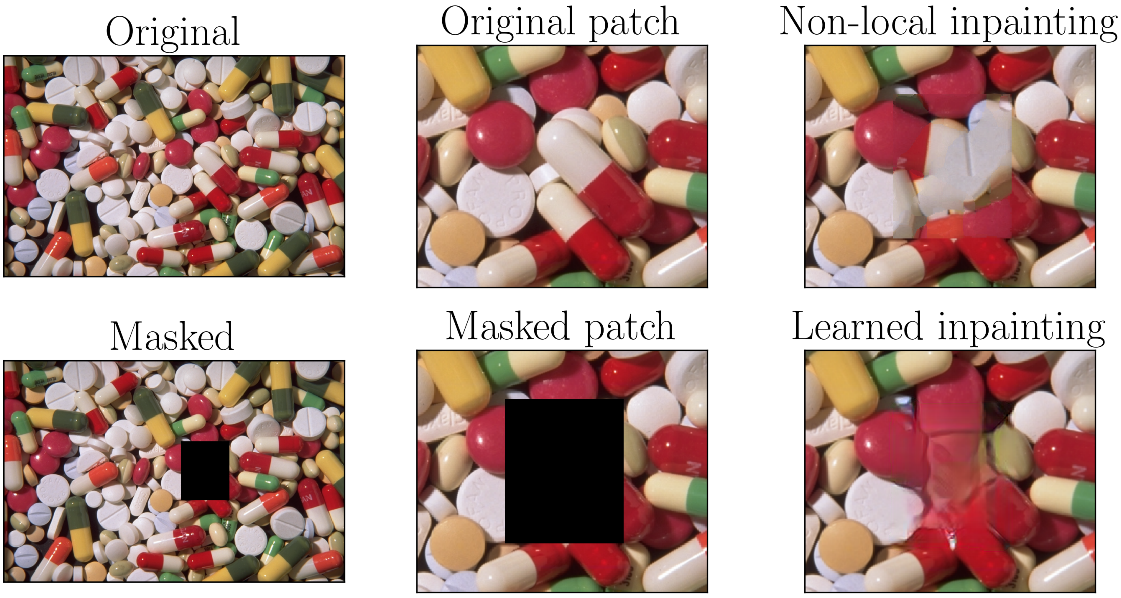


**Figure 1.2** Image from Karel de Gendre[a]. The non-local method uses classical patch-based ideas[b] and the learned result was generated by NVIDIA's deep learning tool[c].

[a] http://profotos.com/pros/profiles/biography.cfm?member=192
[b] http://www.ipol.im/pub/art/2017/189/
[c] https://www.nvidia.com/research/inpainting/.

Splines are a very capable, efficient, and consistent class of interpolators customisable to a given balance of smoothness and locality.

To go beyond this smoothness trade-off, researchers now consider methods which use a large database of images to learn what fine structures occur in natural images. When a new coarse image is presented, a corresponding high resolution image is constructed by combining similar images found in the old database. This task can be accomplished via statistical methods or more current machine learning techniques, examples can be found in Freeman et al. (2002) and Ledig et al. (2017) respectively. In Figure 1.3 we demonstrate the performance of one machine learning tool which is freely available online.

The take-home message of Figure 1.3 is that, in general, the very naive classical method recovers as much information as the modern method, although it is not so visually attractive. The colours are not as accurate and the lines not as sharp but the text is still easily readable in both reconstructions.

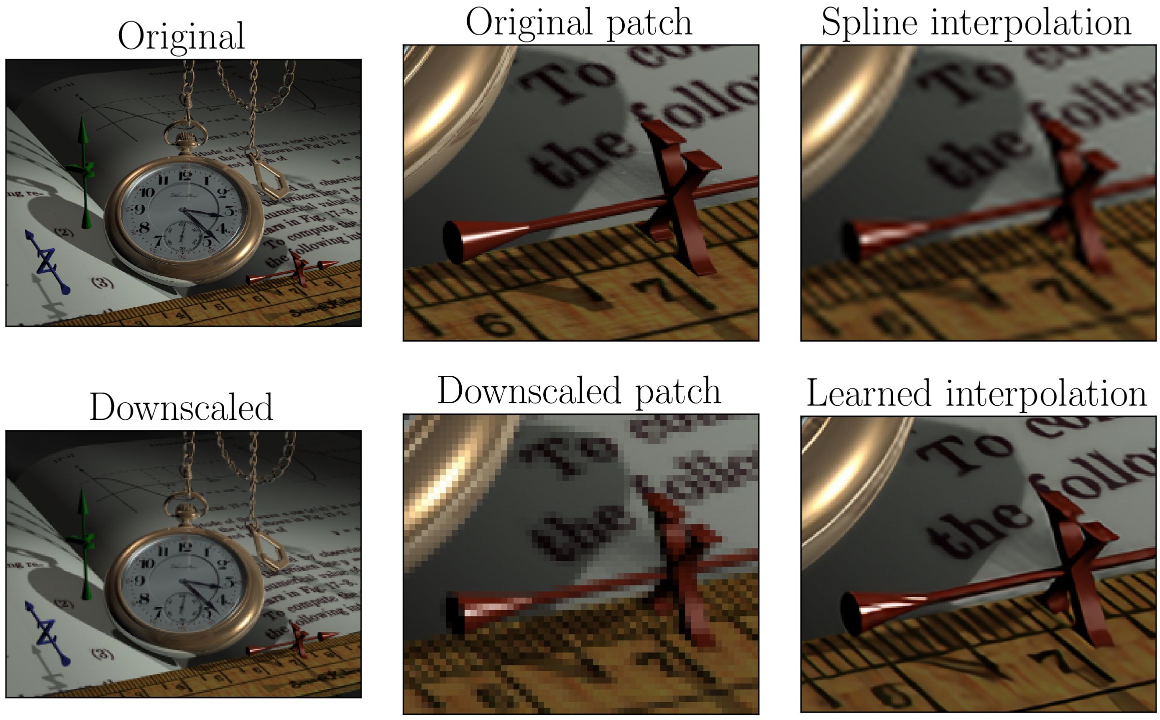


**Figure 1.3** Image from Kevin Odhner[a]. The image is downscaled by a factor of four in both directions and then upscaled again by classical bicubic spline interpolation and an online machine learning based method[b].

[a] jko@home.com
[b] https://imageupscaler.com

### Indirect measurements

Figure 1.4 gives a first flavour of the types of inverse problems important for electron tomography. In particular, F is a sub-sampled X-ray transform which will be defined in detail in Section 1.4.6. Similarly, the inversion method used here is called the *filtered back-projection* which dates back to the first practical applications of this inverse problem in Bracewell (1956) (see Deans, 1983, for more detail). The filtered back-projection will again be introduced formally in Section 1.4.6, here we will simply state the inversion formula:

$$u^* = \operatorname*{argmin}_{u\in L^1([-1,1]^2)} \left\{\|u\|_2 \text{ s. t. } \mathrm{F}(u) = \eta\right\},$$

where F changes to reflect the different available data in each column of Figure 1.4.

The first column of Figure 1.4 shows the exact data when F is the fully sampled X-ray transform and its corresponding reconstruction. In this case $u^*$ is indistinguishable from the ground truth $u^\dagger$. This behaviour is common, if the indirect data is fully sampled at high resolution and free of noise, then even simple reconstruction methods provide accurate reconstructions.

Moving to the second column, adding normally distributed noise to the data and performing the same naive reconstruction results in a 'noisy' reconstruction. This does not completely generalise across all types of indirect measurement, but the key point is that the reconstruction is still a degraded copy of the desired solution with random fluctuations in intensity.

Super-resolution and inpainting can also be interpreted in X-ray inverse problems and both lead to reconstruction artifacts. In the third column, data is sampled coarsely in the $x$ axis and this leads to *streak artifacts*. Similarly, there is a large interval of data missing in the fourth column of Figure 1.4 which leads to *elongation artifacts* in the reconstruction. In each case, reconstruction artifacts appear when the structure of the indirect measurements is not taken into account. These ideas will be made rigorous in Section 1.4.7 and the challenge of overcoming elongation artifacts will be discussed in more depth in Chapter 2.

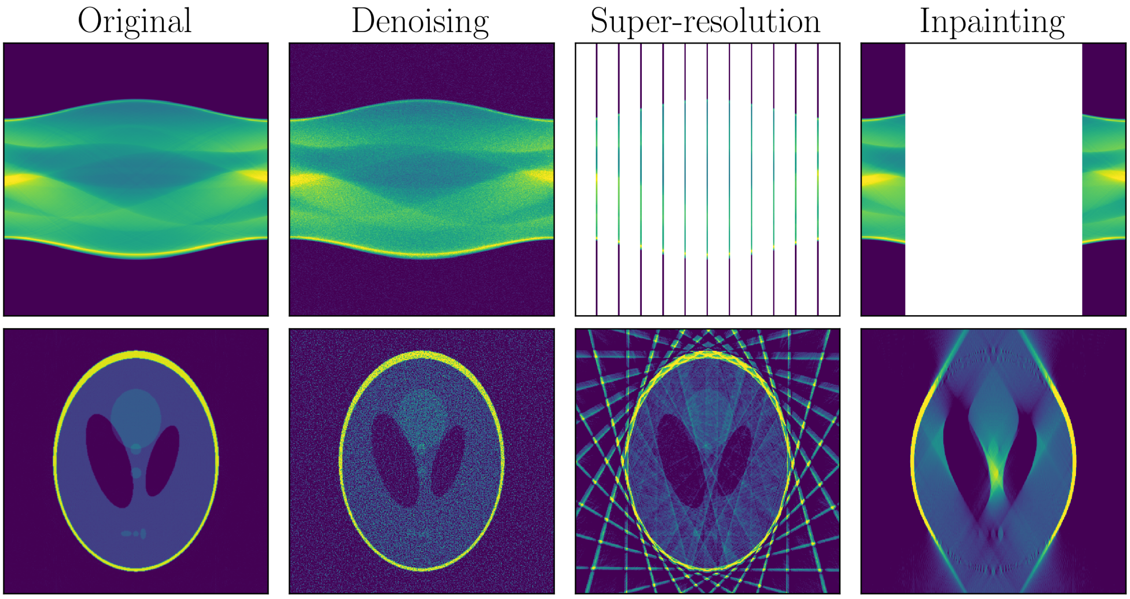


**Figure 1.4** Several naive reconstructions of the modified Shepp-Logan phantom from corrupted indirect measurements. Indirect measurements are given on the top row and corresponding reconstruction on the bottom.

### 1.2.2 Electron microscopy

The main source of inverse problems considered in this work come from electron microscopy. Each of these relies on combining the fundamental properties of electrons with a chosen modality, i.e. the detectors and data sampling pattern of the microscope.

The technology for electron microscopy (EM), namely the ability to manipulate and measure electrons, was first developed at the beginning of the 20th century. The first prototype electron microscope was made in 1931 by Ernst Ruska and Max Knoll and after a succession of refinements, the first commercial microscope was available from 1938 (Freundlich, 1963). In most forms of EM, the idea is that electrons are fired at a sample and any resulting radiation, including the original electrons, are recorded. Every observation contains information which can be used to infer properties of the original sample. The key attraction of using electrons for imaging is that they easy to manipulate and interact very strongly with physical matter (Egerton, 2005). Electrons are controlled by magnets to very high precision and, while very expensive, modern electron microscopes can fit inside a large room.

#### Low-level interactions

We begin by describing the interactions between electrons and atoms, shown in Figure 1.5. At a very coarse level, most electron microscopes work by firing a beam of energy into a sample and measuring some of these different forms of energy which radiate out. The initial electron

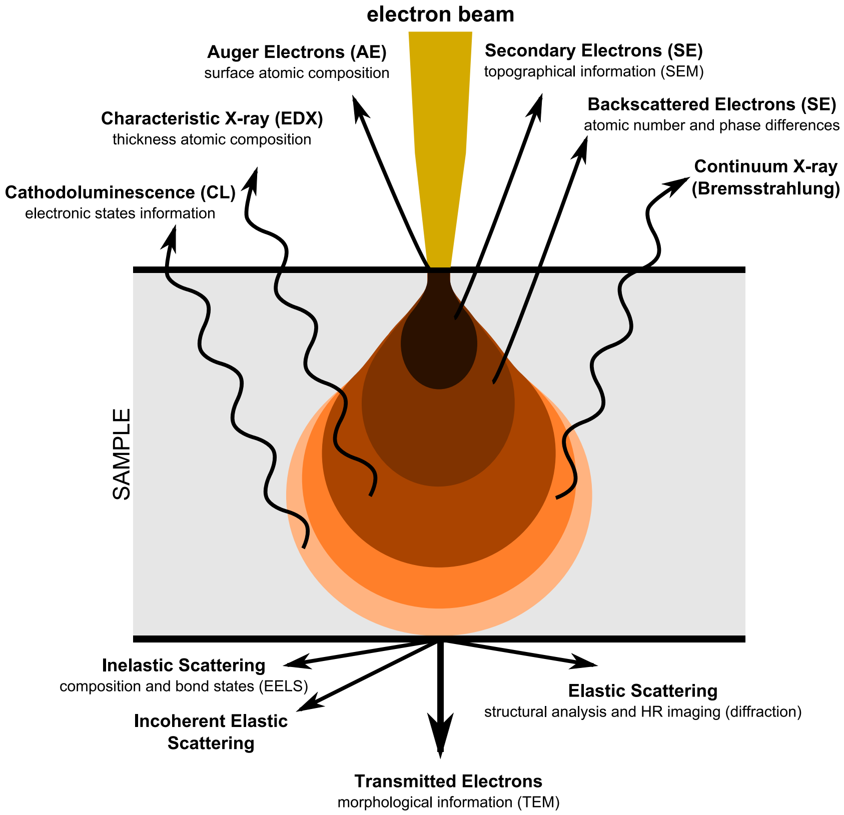

electron beam
Auger Electrons (AE)
surface atomic composition
Secondary Electrons (SE)
topographical information (SEM)
Characteristic X-ray (EDX)
thickness atomic composition
Backscattered Electrons (SE)
atomic number and phase differences
Cathodoluminescence (CL)
electronic states information
Continuum X-ray
(Bremsstrahlung)
SAMPLE
Inelastic Scattering
composition and bond states (EELS)
Elastic Scattering
structural analysis and HR imaging (diffraction)
Incoherent Elastic
Scattering
Transmitted Electrons
morphological information (TEM)

**Figure 1.5** Image from Claudionico-commonswiki demonstrating the interactions of electrons with matter.
https://en.wikipedia.org/wiki/Electron_microscope\#/media/File:Electron_Interaction_with_Matter.svg

beam is typically focused on a very small patch on the top surface of the sample, and then diffuses into the surrounding matter before exiting from the bottom surface. It is not necessary to understand each of these interactions other than to say that each signal encodes physical and chemical information of the sample which can be used to inform a reconstruction. In the course of this work we shall only be interested in the electrons which exit at the bottom of the sample, particularly the transmitted and elastically scattered electrons. The key idea is that heavier samples will scatter more electrons, reducing the number of transmitted electrons which pass straight through the sample. Measuring the number of transmitted electrons indicates how much mass is between the top and bottom surfaces. Counting the number of transmitted vs scattered electrons are two sides of the same coin, if a sample is light/thin then the transmitted number is high (bright field imaging) and the scattered number is low (dark field imaging), c.f. De Rosier and Klug (1968); Leary and Midgley (2019); Midgley and Weyland (2003). For more detailed information, analysing exactly where electrons are scattered to will also allow us to sense more material properties of the sample than simply counting the electrons.

While it is not necessary to look further into Figure 1.5 for this work, it can give some insight into other interesting applications of EM:

- Not all signals are electrons, any squiggly line represents a photon which has been formed during an interaction with the beam and sample
- Not all electrons come from the beam, *secondary electrons* can be released from the sample when the beam interacts with an atom
- Inelastic scattering occurs when an electron from the beam loses energy. That lost energy is converted into signal in the form of one of the arrows pointing upwards.

Inelastic scattering and characteristic X-ray signals form popular modalities for detecting chemical properties of samples, called Electron Energy Loss Spectroscopy (EELS) and Energy Dispersive X-ray (EDX) microscopy respectively. The spectrum of energies of such electrons or X-rays form fingerprints of the structures inside the sample. Each atomic bond has a unique fingerprint and so this tells us which atoms are bonded together, rather than just which are present.

### 1.2.3 Inverse problems from electron microscopes

One electron microscope may support many modalities, each modality has a different acquisition geometry and numerical model. Every combination of microscope detectors and acquisition geometry corresponds to a new inverse problem with distinctive properties.

As was seen in Figure 1.5, each time the electron beam passes through the sample a rich signal is emitted. If the direction and energy of scattered electrons is recorded then this is already a (2+1)D spectral dataset containing information about the sample. Size and resolution of detectors dictates coverage of the sphere of possible scattering angles. Some detectors are also capable of measuring the outgoing electron energy although there are often practical trade-offs. The spectral component of electron imaging can be split into three common cases:

- The detector consists of a single pixel recording 1D energy loss spectra, as is the case for EELS and EDX microscopy.
- The detector records a 2D greyscale *diffraction pattern* covering a portion of the sphere, referred to as *electron diffraction imaging.*
- The detector consists of a single greyscale pixel, as in electron tomography or standard EM.

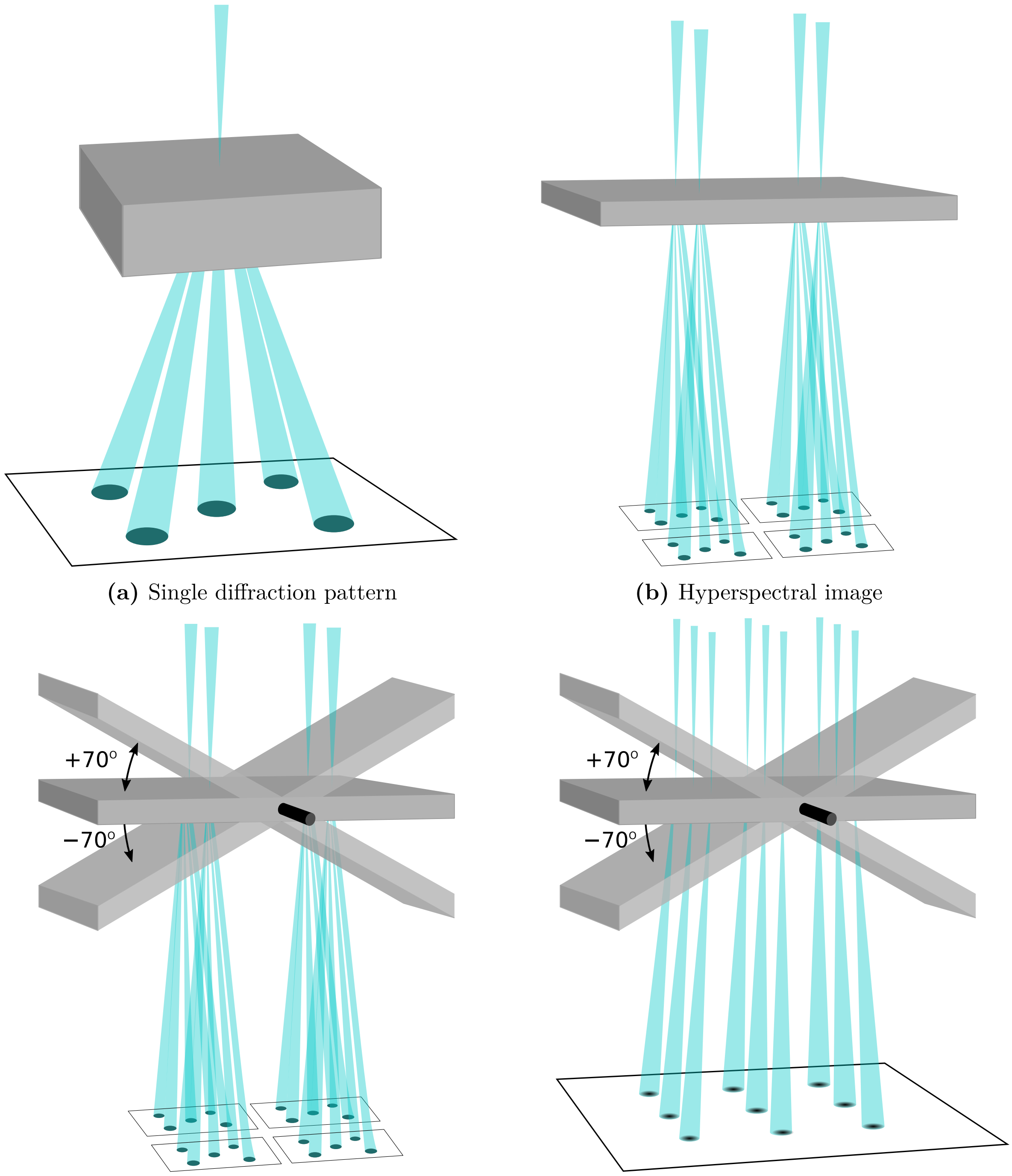
Beyond the detector setup, one single beam often does not provide sufficient information for the desired reconstruction, therefore data is recorded over a sequence of beam positions (and orientations for tomography). Some of the possible scan geometries are depicted in Figure 1.6:

- Figure 1.6a: no scanning. A single electron beam is used to collect (spectral) data.

- Figure 1.6b: the beam is scanned over a grid recording data at each beam position to recover 2D spatial information about a sample.
- Figure 1.6c: scans are performed whilst tilting the sample to different orientations to recover 3D spatial information about a sample.

The greyscale *tilt series* depicted in Figure 1.6d is by far the most common modality for 3D tomographic reconstructions in electron microscopy. In Section 1.4.6 we will show that the corresponding mathematical model is the X-ray transform which allows for the reconstruction of 3D densities from (1+2)D greyscale datasets.

The other scenario focused on in this work is Figure 1.6c where the acquired data is five dimensional, or a 1D tilt series of 2D scans of 2D diffraction patterns. The physical forward model for this can be simulated by classical diffraction methods which will be seen in Section 1.4.3. Both Figures 1.6c and 1.6d are examples of electron tomography modalities. The physics and mathematics are already well established in the scalar case (Leary and Midgley, 2019), although the physical averaging process is currently less well understood in the diffraction case. One model for this purpose is proposed in Chapter 3.



**(a)** Single diffraction pattern

**(b)** Hyperspectral image

**(c)** Hyperspectral tilt series

**(d)** Scalar tilt series

**Figure 1.6** Standard data geometries in EM. Figure 1.6a: every individual electron beam forms a 2D diffraction pattern. Figure 1.6b: scanning the beam over an grid reveals 2D spatial structure of the sample. Figure 1.6c: tilting the specimen reveals 3D spatial structure of the sample. Figure 1.6d: the full spectral data is often not necessary so the beam scan can be viewed as a single greyscale image.

## 1.3 Inverse problems preliminaries

As stated in (1.1), the generic inverse problem is to:

$$\text{find the best} \quad u \quad \text{such that} \quad \mathrm{F}(u) \approx \eta.$$

For the scope of this thesis we will reduce the generality of the formulation. In particular, we assume:

- F is a linear map, which we now denote by $\mathcal{A}\colon \mathbb{U} \to \mathbb{V}$.
- '$\approx$' is in the Euclidean sense in $\mathbb{V}$.
- There exists a *regularisation functional* (or *regulariser*) $\mathrm{g}\colon \mathbb{U} \to \mathbb{R}$ which ranks the quality of $u$ without knowledge of $\eta$. i.e. if $\mathrm{g}(u_1) < \mathrm{g}(u_2)$ then $u_1$ is 'better' than $u_2$.

There are two standard philosophies for reconstruction, optimisation and Bayesian. We will start by describing the motivation of the former, as it is more relevant in this work, before briefly comparing with the Bayesian approach.

Optimisation formulations, or *variational methods* (Smith, 1974; Engl et al., 1996; Benning and Burger, 2018), naturally arise with the desire to find the 'optimal' reconstruction. Combining the above assumptions, we can define a reconstruction method to find $u^*$ such that

$$\mathrm{E}(u^*) = \min_{u\in\mathbb{U}} \mathrm{E}(u) \qquad \text{where} \qquad \mathrm{E}(u) \coloneqq \frac{1}{2}\left\|\mathcal{A}u - \eta\right\|^2 + \mathrm{g}(u). \tag{1.2}$$

Note that if $u^*$ is a minimiser, then both $\|\mathcal{A}u^* - \eta\|$ and $\mathrm{g}(u^*)$ are small, i.e. $\mathcal{A}u^* \approx \eta$ and $u^*$ is the 'best' such solution. If this is a good reconstruction model then this is sufficient to indicate $u^* \approx u^\dagger$.

Almost all of the reconstruction methods discussed in the contributions of this work can be expressed in the form of (1.2). The rest of this section is therefore dedicated to giving some examples of such models, motivating their success, and introducing some numerical methods for solving the corresponding optimisation problems.

The Bayesian approach to inverse problems, see for instance Stuart (2010); Sullivan (2015), provides a wide set of reconstruction methods and theoretical guarantees with distinct motivation to the variational approach. For example, most approximation guarantees will show that $u^* \approx u^\dagger$ with 'high probability' rather than a deterministic bound as in Engl and Grever (1994). For reconstruction methods, there is some overlap as can be derived from Bayes' formula which assigns a probability to each possible reconstruction. This requires the introduction of two probability measures:

- Prior$\colon \mathbb{U} \to \mathbb{R}_{\geq 0}$, the *prior functional* computes the probability of any sample $u$ being the ground truth $u^\dagger$ 'prior' to observing the data. If it is known that $\eta$ is a picture of a house pet, then this immediately suggests that $u^\dagger$ is more likely to be a cat/dog than a lion/wolf.

- $\mathrm{Likelihood}\colon \mathbb{V}^2 \to \mathbb{R}_{\geq 0}$, the *likelihood function* corresponds to the distribution of possible data corruptions. In the case of additive noise with distribution $\mathcal{N}$, this simplifies to $\mathrm{Likelihood}(\eta, \eta') = \mathrm{Likelihood}(\eta - \eta')$ is the probability that $\eta - \eta'$ is a random draw from $\mathcal{N}$ (i.e. Likelihood is the probability density function of $\mathcal{N}$).

Bayes' formula combines these two distributions into the expression:

$$\mathrm{probability}(u = u^\dagger \text{ given } \eta) = \frac{\mathrm{Likelihood}(\mathcal{A}u - \eta)\,\mathrm{Prior}(u)}{\mathrm{probability}(\eta)}.$$

The left-hand side is called the *posterior density* and the term $\mathrm{probability}(\eta)$ is an unknown scaling constant which can typically be ignored.

Following the optimisation mentality, the optimal reconstruction $u^*$ should be the function with highest probability to have formed the data $\eta$ given the prior assumptions on $u^\dagger$. Analytically, this is written as

$$\begin{aligned} u^* &\in \operatorname*{argmax}_{u\in\mathbb{U}} \mathrm{Posterior}(u = u^\dagger \text{ given } \eta) \\ &= \operatorname*{argmin}_{u\in\mathbb{U}} -\log(\mathrm{Posterior}(u = u^\dagger \text{ given } \eta)) \\ &= \operatorname*{argmin}_{u\in\mathbb{U}} -\log(\mathrm{Likelihood}(\mathcal{A}u - \eta)) - \log(\mathrm{Prior}(u)). \end{aligned}$$

This equation is exactly the formulation of (1.2) if the likelihood is chosen for Gaussian white noise and $\mathrm{g}(u) \propto -\log(\mathrm{probability}(u))$.

This argument shows that variational methods can be equivalent to maximising the posterior although for completeness we note that not all functions g correspond to a probability distribution. In particular, the total variational used frequently in this thesis is not consistent with the Bayesian perspective. See Dunbar et al. (2020) for further discussion. Despite the inconsistency, we will adopt the intuitive language of the statistical perspective and refer to $\frac{1}{2}\left\|\mathcal{A}u - \eta\right\|^2$ as a *noise model* and g as a *prior* which encodes the *prior knowledge* of $u^\dagger$.

The Bayesian approach proposes many reconstruction methods based on the posterior distribution, although they will not be used in this work. We have seen that the 'mode' (i.e. maximum) corresponds with the variational method but the 'mean' is a distinct value which can also be used as a reconstruction, for example Latz et al. (2018). A more thorough introduction can be found in Stuart (2010); Sullivan (2015).

### 1.3.1 Normed function spaces

This section introduces the standard notation and results for normed function spaces.

**Definition 1.3.1** (Finite dimensional norms)**.** *For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $p \in [1, \infty)$ we define*

$$\boldsymbol{x} \cdot \boldsymbol{y} := \sum_{i=1}^n x_i y_i, \qquad |\boldsymbol{x}|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}, \qquad \text{and} \qquad |\boldsymbol{x}|_\infty := \max_{i \in [n]} |x_i|.$$

*If not specified, $|\boldsymbol{x}| = |\boldsymbol{x}|_2$. For $p \in [0, \infty]$ we define the normed space*

$$\ell^p(\mathbb{R}^n) = (\mathbb{R}^n, |\cdot|_p)$$

*which is abbreviated to $\ell^p = \ell^p(\mathbb{R}^n)$ when the dimension is clear from surrounding context.*

**Definition 1.3.2** (Infinite dimensional norms)**.** *Let $\mathcal{M}(\Omega)$ be the space of measurable functions mapping from a measurable $\Omega \subset \mathbb{R}^d$ to $\mathbb{R}$. For all $u, v \in \mathcal{M}(\Omega)$ and $p \in [1, \infty)$ we define*

$$\langle u,\ v\rangle := \int_\Omega u(\boldsymbol{x}) v(\boldsymbol{x}) d\boldsymbol{x}, \qquad \|u\|_p := \left(\int_\Omega |u(\boldsymbol{x})|^p d\boldsymbol{x}\right)^{\frac{1}{p}}, \qquad \text{and} \qquad \|u\|_\infty := \sup_{\boldsymbol{x} \in \Omega} |u(\boldsymbol{x})|.$$

*If not specified, $\|u\| = \|u\|_2$. For $p \in [1, \infty]$ we define the normed space*

$$L^p(\Omega) = (\mathcal{M}(\Omega), \|\cdot\|_p)$$

*which is abbreviated to $L^p = L^p(\Omega)$ when the domain is clear from surrounding context.*

**Definition 1.3.3** (Infinite dimensional vector norms)**.** *Let $\mathcal{M}^n(\Omega)$ be the space of measurable functions mapping from a measurable $\Omega \subset \mathbb{R}^d$ to $\mathbb{R}^n$. For all $\vec{u}, \vec{v} \in \mathbb{U} \subset \mathcal{M}^n(\Omega)$ and $p \in [1, \infty)$, $q \in [1, \infty]$ we define*

$$\langle \vec{u},\ \vec{v}\rangle := \int_\Omega \vec{u}(\boldsymbol{x}) \cdot \vec{v}(\boldsymbol{x}) d\boldsymbol{x}, \qquad \|\vec{u}\|_{p,q} := \||\vec{u}|_q\|_p := \left(\int_\Omega |\vec{u}(\boldsymbol{x})|_q^p d\boldsymbol{x}\right)^{\frac{1}{p}},$$

$$\text{and} \qquad \|\vec{u}\|_{\infty,q} := \sup_{\boldsymbol{x} \in \Omega} |\vec{u}(\boldsymbol{x})|_q.$$

*For $p \in [1, \infty]$ we define the normed space*

$$L^{p,q}(\Omega, \mathbb{R}^n) = (\mathcal{M}^n(\Omega), \|\cdot\|_{p,q})$$

*which is abbreviated to $L^{p,q} = L^{p,q}(\Omega)$ when clear from surrounding context. If $q$ is not specified then assume $q = 2$. If neither $p$ or $q$ are specified then assume $\|u\| = \|u\|_{2,2}$.*

**Definition 1.3.4** (Linear operator norms)**.** *Suppose $\mathcal{A}\colon \mathbb{U} \to \mathbb{V}$ is a linear operator for some (finite or infinite) dimensional spaces $\mathbb{U}$ and $\mathbb{V}$. We define*

$$\|\mathcal{A}\|_{\mathbb{U} \to \mathbb{V}} := \sup_{u \in \mathbb{U}} \frac{\|\mathcal{A}u\|_{\mathbb{V}}}{\|u\|_{\mathbb{U}}}.$$

*If, for instance,* $\mathbb{U} = L^p$ *and* $\mathbb{V} = L^q$ *then we abbreviate this to* $\|\mathcal{A}\|_{p,q} = \|\mathcal{A}\|_{L^p \to L^q}$. *If* $p = q$ *then this is further abbreviated to* $\|\mathcal{A}\|_p = \|\mathcal{A}\|_{p,p}$.

**Definition 1.3.5** (Smooth norms and semi-norms)**.** *Let* $C^k(\Omega)$ *denote the space of functions on an open domain* $\Omega \subset \mathbb{R}^d$ *with* $k$ *continuous derivatives. For* $u \in C^k(\Omega)$ *we define the semi-norm*

$$|u|_{C^k} = \sup_{\boldsymbol{x} \in \Omega} \left\| \nabla^k u(\boldsymbol{x}) \right\|_2$$

*and the full norm*

$$\|u\|_{C^k} = \|u\|_\infty + |u|_{C^k}.$$

*If* $\mathcal{A} \colon \mathbb{U} \to C^k$ *then the corresponding norms are*

$$|\mathcal{A}|_{\mathbb{U} \to C^k} = \sup_{u \in \mathbb{U}} \frac{|\mathcal{A}u|_{C^k}}{\|u\|_{\mathbb{U}}}, \qquad \|\mathcal{A}\|_{\mathbb{U} \to C^k} = \sup_{u \in \mathbb{U}} \frac{\|\mathcal{A}u\|_{C^k}}{\|u\|_{\mathbb{U}}}.$$

**Definition 1.3.6** (Duality)**.** *For a normed space* $\mathbb{U}$ *we denote its dual space*

$$\mathbb{U}^* = \{\mathcal{A} \colon \mathbb{U} \to \mathbb{R} \text{ s.t. } \mathcal{A} \textit{ is linear and } \|\mathcal{A}\| < \infty\}.$$

**Theorem 1.3.7** (Riesz representation theorem)**.** *Suppose* $\mathbb{U} = L^p(\Omega)$ *and* $\mathbb{V} = L^q(\Omega)$ *for* $p, q \in (1, \infty)$ *and* $\frac{1}{p} + \frac{1}{q} = 1$. *We can write* $\mathbb{U}^* = \mathbb{V}$ *under the understanding*

$$v(u) \coloneqq \langle u,\ v \rangle .$$

*This is an isometric embedding, i.e.*

$$|\langle u,\ v\rangle| \leq \|u\|_p \|v\|_q$$

*and*

$$\|u\|_p = \sup_{\|v\|_q \leq 1} \langle u,\ v\rangle .$$

*If* $p = 1$ *and* $q = \infty$ *then the same statement holds.*

*In the finite (or countable) dimensional setting of* $\mathbb{U} = \ell^p(\mathbb{R}^n)$ *etc. the statements are equivalent.*

**Definition 1.3.8** (Notions of convergence)**.** *Let* $\mathbb{U}$ *be a normed space with dual space* $\mathbb{V} = \mathbb{U}^*$, $(u_i)_{i=1}^\infty$ *a sequence in* $\mathbb{U}$, *and* $u \in \mathbb{U}$. *We say* $u_i$ *converges to* $u$, *or* $u_i \to u$, *if*

$$\|u_i - u\|_{\mathbb{U}} \to 0.$$

*We say* $u_i$ *converges to* $u$ *weakly, or* $u_i \rightharpoonup u$, *if*

$$\langle u_i - u,\ v\rangle \to 0 \qquad \textit{for all } v \in \mathbb{V}.$$

*We say a sequence* $(v_i)_{i=1}^\infty \subset \mathbb{V}$ *converges to* $v \in \mathbb{V}$ *weakly-*∗*, or* $v_i \overset{*}{\rightharpoonup} v$*, if*

$$\langle u,\ v_i - v\rangle \to 0 \qquad \text{for all } u \in \mathbb{U}.$$

*Note that in most cases that* $\|\cdot\|_{\mathbb{U}} = \|\cdot\|_p$ *with* $p < \infty$*, weak and weak-*∗ *convergence are equivalent. The only exception is when* $p = 1$ *and* $\dim(\mathbb{U}) = \infty$*, then weak-*∗ *is strictly weaker than weak convergence.*

$$\text{If } \mathbb{U} = L^1(\Omega),\ \operatorname{diam}(\Omega) < \infty, \text{ then} \qquad u_i \overset{*}{\rightharpoonup} u \iff \langle u_i - u,\ v\rangle \to 0 \text{ for all } v \in C^0(\Omega).$$
$$\text{If } \mathbb{U} = \ell^1(\mathbb{R}^\infty), \text{ then} \qquad u_i \overset{*}{\rightharpoonup} u \iff \langle u_i - u,\ v\rangle \to 0 \text{ for all } v \in \ell^\infty(\mathbb{R}^\infty) \text{ s.t. } |v_j| \to 0.$$

**Theorem 1.3.9** (Banach-Alaoglu compactness theorem)**.** *Let* $\mathbb{U}$ *be a normed space and* $(u_i)_{i=1}^\infty$ *a bounded sequence in* $\mathbb{U}$*, i.e.* $\sup_{i\in\mathbb{N}} \|u_i\|_{\mathbb{U}} < \infty$*. There exists a subsequence* $n_i$ *such that* $n_1 < n_2 < \ldots$ *and*

$$u_{n_i} \text{ converges in } \mathbb{U}\ldots \quad \begin{cases} \text{strongly} & \text{if } \dim(\mathbb{U}) < \infty \\ \text{weakly} & \text{if } \mathbb{U} \subset L^p(\Omega) \text{ or } \ell^p(\mathbb{R}^\infty),\ p \in (1,\infty) \\ \text{weak-}* & \text{else.} \end{cases}$$

### 1.3.2 Common properties of variational functionals

When developing a variational method in inverse problems, it is very common to ensure that E from (1.2) is a *proper*, *convex* and *weakly lower-semicontinuous* functional with bounded sublevel sets. While not essential, this set of properties comfortably guarantees that there exist minimisers $u^*$ which are well-defined and efficiently computable. In this section we will define each of these terms and explore its usefulness in the context of inverse problems.

We start with the most basic properties.

**Definition 1.3.10.** *Let* $\mathrm{E}\colon \mathbb{U} \to \overline{\mathbb{R}}$ *be an arbitrary function.*

- *If* $\min\limits_{u\in\mathbb{U}} \mathrm{E}(u) < \infty$*, then* E *is called* proper*.*
- *If* $\operatorname{diam}(\{u \in \mathbb{U} \text{ s.t. } \mathrm{E}(u) < t\}) < \infty$ *for every* $t > \min_{u\in\mathbb{U}} \mathrm{E}(u)$*, then* E *is said to have* bounded sublevel sets*.*
- *If* $\mathrm{E}(u) = \min\limits_{u' \rightharpoonup u} \mathrm{E}(u')$ *for all* $u \in \mathbb{U}$*, then* E *is called* weakly lower-semicontinuous*.*

As previously stated, these properties are not strictly necessary, however provide relatively relaxed conditions which guarantee that E makes sense as a variational model. If E is not proper, then $\mathrm{E} = \infty$ almost everywhere and so has no minimisers. The functional $\mathrm{E}(u) = e^{-u}$ is a simple example of a convex function which has unbounded sublevel sets and the minimiser is $u = +\infty$. Infinity can't be placed in a microscope so it is natural to avoid this 'minima at infinity' eventuality. Semicontinuity is also a very natural stability assumption in optimisation which guarantees that minimisers exist at every minima of E. A simple example where this goes wrong is

$$\mathrm{E}(u) = \begin{cases} \infty & u \leq 0 \\ u^2 & u > 0 \end{cases}.$$

We might say that the minimiser is clearly $u = 0$ but with the conflict that $\mathrm{E}(0) \neq \min_u \mathrm{E}(u)$, indeed we have $\mathrm{E}(0) = \infty$. This can happen even when E is (strongly) convex, and so lower semicontinuity is assumed to exclude these cases.

Already with these three assumptions we can start to see the skeleton of the argument for finding reconstructions with variational methods. The following lemma shows that any 'good' sequence of approximate minimisers must also approximate $u^*$.

**Lemma 1.3.11.** *Suppose* E *is a proper, lower-semicontinuous function with bounded level sets and* $(u_i)_{i=1}^\infty$ *is a sequence such that* $\inf_i \mathrm{E}(u_i) = \inf_{u\in\mathbb{U}} \mathrm{E}(u)$. *There exists* $u^* \in \operatorname{argmin}_{u\in\mathbb{U}} \mathrm{E}(u)$ *and a subsequence* $n_i$ *such that* $u_{n_i} \to u^*$ *where convergence is in the topology described by Theorem 1.3.9.*

*Proof.* If $\mathrm{E}(u_i) = \inf_{u\in\mathbb{U}} \mathrm{E}(u)$ for some $i$, then we can take $u^* = u_i$. Otherwise, there exists a monotonically decreasing subsequence $\mathrm{E}(u_{n_i}) \searrow \inf_i \mathrm{E}(u_i) = \inf_{u\in\mathbb{U}} \mathrm{E}(u)$. This subsequence lies in a (bounded) sublevel set to which can be applied Theorem 1.3.9 as required. □

The take-home of this lemma is that it is typically very easy to find a sequence with $\mathrm{E}(u_i) \to \inf_i \mathrm{E}(u_i)$, which immediately (implicitly) identifies a minimiser as a weak limit of a subsequence. There are a few standard techniques for upgrading from weak to strong convergence, for example:

- If $\dim(\mathbb{U}) < \infty$, then weak and strong convergence are equivalent.
- If E is strongly convex, then $\mathrm{E}(u) \approx \min \mathrm{E} \implies \|u - u^*\| \approx 0$.
- If $u_i$ are bounded in a stronger topology, for instance $W^{1,p}$, then Rellich's embedding theorem guarantees that a subsequence converges strongly in $L^q$ for a range of $q$.

Finite dimensionality is often only true at the numerical level and many functions in applications are not strongly convex. Rellich's theorem is often applicable when the regularisation functional includes derivatives.

A more generalised notion of strong convexity is the *Kurdyka-Łojasiewicz* property which can also be leveraged for stronger convergence guarantees in many finite dimensional convex optimisation problems (Bolte et al., 2007). In particular, one can often show that $u_i$ converge strongly (not just on a subsequence) at a guaranteed rate (Attouch et al., 2010).

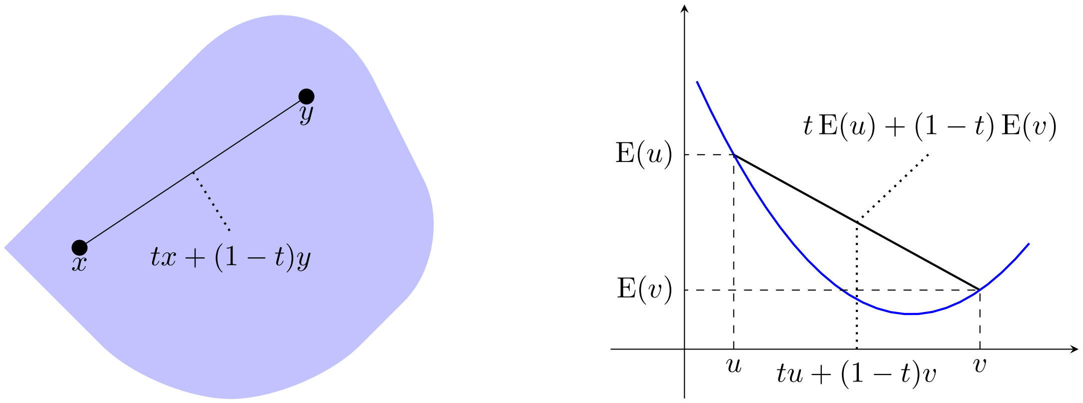


**Figure 1.7** Geometrical definition of a convex set (left) and a convex function (right)

### Convexity

The final classical assumption on E in variational methods center around different forms of convexity. The textbook definition is as follows.

**Definition 1.3.12.** *A set* $\mathbb{D} \subset \mathbb{R}^d$ *is called* convex *if*

$$tx + (1-t)y \in \mathbb{D} \qquad \text{for all } x, y \in \mathbb{D},\ t \in [0,1].$$

*A function* $\mathrm{E}\colon \mathbb{U} \to \overline{\mathbb{R}}$ *is called* convex *if*

$$\mathrm{E}(tx + (1-t)y) \leq t\,\mathrm{E}(x) + (1-t)\,\mathrm{E}(y) \qquad \text{for all } x, y \in \mathbb{D},\ t \in [0,1].$$

E *is called* $\mu$-strongly convex *for* $\mu \geq 0$ *if*

$$\mathrm{E}(tx + (1-t)y) + \mu\tfrac{t(1-t)}{2}\,\|x-y\|_2^2 \leq t\,\mathrm{E}(x) + (1-t)\,\mathrm{E}(y) \qquad \text{for all } x, y \in \mathbb{D},\ t \in [0,1].$$

These are very geometrical definitions which are also sketched in Figure 1.7. The first advantage of convexity is that it guarantees some level of uniqueness of minimisers.

**Lemma 1.3.13** ((Boyd et al., 2004, Section 3.1.6))**.** *If* $\mathrm{E}\colon \mathbb{U} \to \overline{\mathbb{R}}$ *is proper, convex, lower-semicontinuous and has bounded sublevel sets, then*

$$C := \{u \in \mathbb{U} \text{ s.t. } \mathrm{E}(u) = \min_{u' \in \mathbb{U}} \mathrm{E}(u')\}$$

*is a closed, convex, non-empty set. If* E *is* $\mu$*-strongly convex with* $\mu > 0$*, then* $C$ *is a single point.*

### Convex differentiability

Differentiability is important in optimisation because they provide an alternative characterisation of minimisers and, as will be shown in Section 1.3.4, are integral to the construction of many numerical methods. Convexity guarantees a very practical notion of derivatives.

**Definition 1.3.14.** *Let* $\mathrm{E}\colon \mathbb{U}\to\overline{\mathbb{R}}$ *and fix* $u\in\mathbb{U}$*. We say that* $\mathrm{E}$ *is* (Fréchet) differentiable *at* $u$ *with derivative* $\nabla\mathrm{E}(u)\in\mathbb{U}^*$ *if*

$$\limsup_{v\to u}\frac{|\,\mathrm{E}(u)+\langle\nabla\mathrm{E}(u),\ v-u\rangle-\mathrm{E}(v)|}{\|v-u\|_{\mathbb{U}}}=0.$$

*If* $\mathrm{E}$ *is a convex function, then we define the* subdifferential $\partial\,\mathrm{E}(u)$ *to be the set of dual elements* $\varphi\in\mathbb{U}^*$ *such that*

$$\limsup_{v\to u}\frac{\mathrm{E}(u)+\langle\varphi,\ v-u\rangle-\mathrm{E}(v)}{\|v-u\|_{\mathbb{U}}}\leq 0.$$

*We say that* $\varphi$ *is a subderivative (or subgradient) of* $\mathrm{E}$ *at* $u$*.*

**Lemma 1.3.15** ((Clarke, 1990, Proposition 2.2.7))**.** *If* $\mathrm{E}\colon\mathbb{U}\to\overline{\mathbb{R}}$ *is convex, then for each* $u\in\mathbb{U}$

$$\varphi\in\partial\,\mathrm{E}(u)\qquad\iff\qquad \mathrm{E}(v)\geq\mathrm{E}(u)+\langle\varphi,\ v-u\rangle\qquad \textit{for all } v\in\mathbb{U}.$$

*In particular,*

$$u\in\operatorname{argmin}\mathrm{E}\iff 0\in\partial\,\mathrm{E}(u).$$

An immediate consequence of this lemma is confirmation that convex functions have no bad local minima or critical points. If $u$ is a local minima, then it must also be a global minima. A slightly stronger inference is that local gradients of convex functions give global information. This strong link between local and global is key to why convexity is valuable for developing numerical and analytical results in inverse problems.

### Convex duality

Convex duality is a generalisation of the Legendre transform which pairs every convex function to another unique partner. This relationship can be utilised to reveal better analytical or numerical structures.

**Definition 1.3.16.** *If* $\mathrm{E}\colon\mathbb{U}\to\overline{\mathbb{R}}$ *is convex, then we define its* convex conjugate *(or Fenchel dual) to be*

$$\mathrm{E}^*\colon\mathbb{U}^*\to\overline{\mathbb{R}},\qquad \mathrm{E}^*(\varphi)\coloneqq\sup_{u\in\mathbb{U}}\langle\varphi,\ u\rangle-\mathrm{E}(u).$$

**Theorem 1.3.17** ((Boyd et al., 2004, Section 3.3))**.** *If* $\mathrm{E}\colon\mathbb{U}\to\overline{\mathbb{R}}$ *is proper, convex and lower-semicontinuous, then*

- $\mathrm{E}^*\colon\mathbb{U}\to\overline{\mathbb{R}}$ *is proper, convex and lower-semicontinuous,*

- $(\mathrm{E}^*)^* = \mathrm{E}$,
- *and, for any* $u \in \mathbb{U}$, $\varphi \in \mathbb{U}^*$

$$\mathrm{E}(u) + \mathrm{E}^*(\varphi) \leq \langle \varphi,\ u \rangle .$$

*Furthermore, the following are equivalent:*

- $\varphi \in \partial\, \mathrm{E}(u)$,
- $u \in \partial\, \mathrm{E}^*(\varphi)$,
- *and* $\mathrm{E}(u) + \mathrm{E}^*(\varphi) = \langle \varphi,\ u \rangle$.

In general it is difficult to compute dual functions explicitly however, there are several simple examples. If $p \in [1, \infty]$ and $\frac{1}{p} + \frac{1}{p^*} = 1$, then

$$\begin{array}{rll} \mathrm{E}(u) = \|u\|_p & \mathrm{E}^*(\varphi) = \begin{cases} 0 & \|\varphi\|_{p^*} \leq 1 \\ \infty & \text{else} \end{cases}, & \\ \mathrm{E}(u) = \frac{1}{p}\|u\|_p^p & \mathrm{E}^*(\varphi) = \frac{1}{p^*}\|\varphi\|_{p^*}^{p^*} & p \notin \{1, \infty\}, \\ \text{and } \ \mathrm{E}(u) = \mathrm{F}(\mathcal{A}u + b) & \mathrm{E}^*(\mathcal{A}^*\varphi) = \mathrm{F}^*(\varphi) - \langle b,\ \varphi \rangle & \end{array}$$

for any proper, convex and lower-semicontinuous F. Note that each of these examples is reflexive in the sense that $\mathrm{E}^{**} = \mathrm{E}$.

### 1.3.3 Examples of variational functionals

In (1.2) we proposed the general variational formulation

$$\mathrm{E}(u) = \frac{1}{2}\|\mathcal{A}u - \eta\|^2 + \mathrm{g}(u).$$

This section introduces key examples of regularisers g which cover the most relevant literature for this thesis.

#### Tikhonov regularisation

The most classical reconstruction methods in inverse problems focus around regularisation using the Euclidean norm. The canonical example of regularisation, called Tikhonov regularisation, takes $\mathrm{g}(u) = \frac{\alpha}{2}\|u\|_2^2$ for some $\alpha \geq 0$ leading to

$$u^* = \operatorname{argmin} \frac{1}{2}\|\mathcal{A}u - \eta\|^2 + \frac{\alpha}{2}\|u\|_2^2 = (\mathcal{A}^*\mathcal{A} + \alpha)^{-1}\mathcal{A}^*\eta.$$

With $\alpha = 0$, this is called the least-squares solution with the Moore-Penrose pseudoinverse $\mathcal{A}^\dagger = (\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*$ on the right-hand side. The corresponding optimisation formulation is

$$u^* = \operatorname{argmin}\left\{\|u\|_2 \text{ s. t. } \mathcal{A}^*\mathcal{A}u = \mathcal{A}^*\eta\right\}$$

where $\mathcal{A}^*\mathcal{A}u = \mathcal{A}^*\eta$ is called the *normal equation* for least squares optimisation. This is very computationally efficient and performs reasonably well in scenarios where the data is of sufficiently high quality. The reconstruction method in Figure 1.4 is exactly a discretised version of the $\alpha = 0$ case.

Writing $\mathrm{g}(u) = \frac{1}{2}\|\sqrt{\alpha}u\|_2^2$ and replacing $\alpha$ with a general linear operator generalises the regularisation to a broad class of linear reconstruction methods. While linear methods are often very efficient and easy to analyse, they typically do not give state-of-the-art performance without further manipulation. There is a current push in machine learning to replace linear with piecewise linear with much improved performance. This topic is reviewed by Unser (2019); Arridge et al. (2019) and a specific example in electron microscopy is given by Pelt and Batenburg (2013).

### Sparse regularisation

The philosophy of inverse problems is that any object can be reconstructed from sufficiently good data, relative to the complexity of the object. 'Simple' objects can be reconstructed from 'worse' data. In the case of Tikhonov regularisation, 'simple' was defined as small in Euclidean norm but different norms modify the regularising effect in different ways. A powerful alternative has been the idea of sparsity, asserting that $u^\dagger(x) \neq 0$ for only a small number of points $x$, interpreted either for continuous indices $x \in \mathbb{R}^d$ or countable $x \in \mathbb{N}$. The landmark results of compressed sensing (Donoho, 2006; Candès et al., 2006) show that such a sparsity regularisation can be achieved exactly for some choices of $\mathcal{A}$ and $\mathrm{g}(u) = \alpha\|u\|_1$ for some $\alpha \geq 0$. This idea dates back at least another 30 years to the works of Claerbout and Muir (1973); Högbom (1974) and can be demonstrated by a simple example. For any $b \in \mathbb{R}$ and $\alpha \geq 0$ observe that:

$$\begin{aligned}
\mathrm{E}(u) &= \frac{1}{2}|u-b|^2 + \frac{\alpha}{2}|u|^2 & &\implies & u^* &= \frac{b}{1+\alpha} \\
\mathrm{E}(u) &= \frac{1}{2}|u-b|^2 + \alpha|u| & &\implies & u^* &= \begin{cases} b - \alpha\,\mathrm{sign}(b) & 0 \leq \alpha < |b| \\ 0 & \alpha \geq |b| \end{cases} .
\end{aligned}$$

In Tikhonov regularisation, $u^*$ is never exactly 0 but with the $L^1$-norm this is achieved whenever $\alpha \geq |b|$. A more detailed motivation is given by Unser et al. (2016); Boyer et al. (2019) where they prove the following explicit reconstruction characterisation.

**Theorem 1.3.18** ((Boyer et al., 2019, Proposition 5))**.** *Let $\mathbb{U}$ be a subspace of $\mathcal{M}(\mathbb{R}^d)$. Suppose $\mathrm{E}\colon \mathbb{U} \to \mathbb{R}$ is defined as*

$$\mathrm{E}(u) = \frac{1}{2} \left\|\mathcal{A}u - \eta\right\|^2 + \alpha \left\|\mathcal{L}u\right\|_1$$

*for some $\alpha \geq 0$ and surjective linear operator $\mathcal{L}\colon \mathbb{U} \to \mathcal{M}(\Omega)$.*

*There exists a minimiser $u^*$ of the form*

$$u^* = \sum_{i=1}^n a_i \mathcal{L}^\dagger \delta_{\boldsymbol{x}_i} + u_{ker} \qquad \text{for some } n \leq \dim(\eta),\ a_i \in \mathbb{R},\ \boldsymbol{x}_i \in \Omega,\ \mathcal{L}u_{ker} = 0$$

*where $\mathcal{L}^\dagger$ is the standard pseudo-inverse and $\delta_{\boldsymbol{x}}$ the Dirac function at $\boldsymbol{x}$. Furthermore, $\left\|\mathcal{L}u\right\|_1 = \left\|\boldsymbol{a}\right\|_1 = \sum_{i=1}^n |a_i|$.*

This theorem exactly characterises what ‘simple’ means in the context of $L^1$-norm penalisation. If the dimension of $\eta$ is $n$, then $\mathcal{L}u^*$ must be non-zero at at most $n$ points. These points $\boldsymbol{x}_i$ and weights $a_i$ are then chosen to balance fitting the data, $\left\|\mathcal{A}u - \eta\right\|^2$, and being small in the $L^1$ sense, $\left\|\boldsymbol{a}\right\|_1$.

There are some applications where the inclusion of $\mathcal{L}$ is unnecessary, for instance images of stars in the night sky are intrinsically sparse, however, choice of $\mathcal{L}$ has allowed these methods to be applied to any natural image. The classical choice for $\mathcal{L}$ is to be a wavelet basis which takes advantage of the piecewise smooth nature of real world objects, for example in Chan et al. (2003); Lustig et al. (2007). In the case of wavelets, $\mathcal{L}$ is an orthogonal matrix, however, in other cases the full generality is necessary. Curvelet and sheerlet frames aim to emphasise the curvature found in natural images and result in non-orthogonal dictionaries (Candes et al., 2006; Kutyniok and Labate, 2012). Outside of inverse problems, thing-lets have also been hugely successful in many areas of applied mathematics, such as in the JPEG2000 standard for general image and video compression (Unser and Blu, 2003). They are numerically very simple and fast whilst being exceptionally efficient at representing general natural signals.

### Total variation

Total variation regularisation was first proposed in the seminal work of Rudin et al. (1992) and fits within the same ideology of sparsity but deserves particular mention as it is very common in applications, including this thesis, and requires a modification of Theorem 1.3.18.

**Definition 1.3.19.** *The* total variation *functional* $\mathrm{TV}\colon \mathcal{M}(\Omega) \to \overline{\mathbb{R}}$ *is defined as*

$$\mathrm{TV}(u) \coloneqq \sup_{\varphi \in C_c^1(\Omega)} \left\{ \langle u,\ \operatorname{div} \varphi \rangle \ \text{s.t.}\ \left\|\varphi\right\|_{\infty,2} \leq 1 \right\}. \tag{1.3}$$

*The space of* bounded variation *is denoted*

$$\mathbb{BV}(\Omega) \coloneqq \left\{ u \in \mathcal{M}(\Omega) \ \text{s.t.}\ \left\|u\right\|_1 + \mathrm{TV}(u) < \infty \right\}.$$

*This follows the constructions of Ziemer (1989); Ambrosio et al. (2000).*

Comments on the natural extension of total variation to colour images can be found in Chan et al. (2001); Ehrhardt and Arridge (2013) although will not be relevant to the scope of this work.

For functions $u \in W^{1,1}(\Omega)$, total variation becomes $\mathrm{TV}(u) = \|\nabla u\|_{1,2}$ where $\nabla u$ is the weak derivative of $u$. In general, $\mathbb{BV} \supset W^{1,1}$ is the set of functions such that $|\nabla u|$ can be understood in a weak sense. This sometimes leads to the notation $\mathrm{TV}(u) = \||\mathrm{D}u|\|_1$ where $\mathrm{D}u$ is to be understood as a distributional derivative.

With this heuristic understanding, if $\mathrm{g}(u) = \alpha\,\mathrm{TV}(u)$, then this regulariser promotes sparse gradients in $u^*$. In other words, $u^*$ should be piecewise constant. This is a mantra which often works surprisingly well in practice and is formalised in the following theorem.

**Theorem 1.3.20** ((Boyer et al., 2019, Theorem 2))**.** *Suppose* $\mathrm{E}\colon \mathbb{U} \to \mathbb{R}$ *is defined as*

$$\mathrm{E}(u) = \frac{1}{2}\|\mathcal{A}u - \eta\|^2 + \alpha\,\mathrm{TV}(u)$$

*for some* $\alpha \geq 0$. *There exists a minimiser* $u^* \in \mathbb{BV}(\Omega)$ *of the form*

$$u^* = \sum_{i=1}^n a_i \mathbb{1}_{\omega_i} + c \qquad \textit{for some } n \leq \dim(\eta),\ a_i, c \in \mathbb{R},\ \omega_i \subset \Omega$$

*such that* $\omega_i$ *are 'simple sets'. Full definition of this is given by Ambrosio et al. (2001) but, informally, they are simply connected sets with no holes and sufficiently smooth boundaries.*

This confirms that $u^*$ is piecewise constant and has some form of nice levelsets.

Analytically, $\mathbb{BV}$ is used because it is a large space which allows for discontinuous objects but is small enough to guarantee some regularity. Discontinuous functions are necessary in imaging as they allow the representation of edges, even this document is discontinuous as it jumps from white page to black text. Weak derivatives are not sufficient for this task, even the 1D Heaviside function only has a distributional derivative:

$$u(x) = \begin{cases} 1 & |x| \leq 1 \\ 0 & \text{else} \end{cases} \quad \text{for all } x \in \mathbb{R} \qquad \implies \qquad \nabla u = \delta_{-1} - \delta_1.$$

Therefore, for every $p \in [1, \infty]$, $u \in L^p(\mathbb{R})$ but $\nabla u \in \mathcal{M}(\mathbb{R}) \setminus L^p(\mathbb{R})$. This is sufficient to show that Sobolev spaces (with non-negative smoothness) are too small to contain discontinuous functions however the BV space is not. On the other hand, the following theorem shows that all BV functions posses some smoothness properties.

**Theorem 1.3.21** ((results from Ziemer, 1989)). *Fix $u \in \mathbb{BV}(\Omega)$ for some bounded $\Omega \subset \mathbb{R}^d$, let*

$$u^-(\boldsymbol{x}) \coloneqq \liminf_{r\to 0} \fint_{|\boldsymbol{y}-\boldsymbol{x}|\leq r} u(\boldsymbol{y}),$$
$$u^+(\boldsymbol{x}) \coloneqq \limsup_{r\to 0} \fint_{|\boldsymbol{y}-\boldsymbol{x}|\leq r} u(\boldsymbol{y}),$$

*and define the edge-set*

$$\mathbb{E} \coloneqq \left\{\boldsymbol{x} \in \Omega \text{ s. t. } u^-(\boldsymbol{x}) \neq u^+(\boldsymbol{x})\right\}.$$

*The following statements hold:*

- $\|u\|_{\frac{d}{d-1}} \lesssim_\Omega \mathrm{TV}(u)$*. If $d = 1$ then $u \in L^\infty$, or if $d = 2$ then $u \in L^2$.*
- *$u(\boldsymbol{x}) = u^-(\boldsymbol{x}) = u^+(\boldsymbol{x})$ is continuous for $\mathbb{R}^d$-almost every $\boldsymbol{x}$*
- *For $\mathbb{R}^{d-1}$-almost every $\boldsymbol{x} \in \mathbb{E}$, if $\boldsymbol{\nu}$ is the oriented normal to $\mathbb{E}$ at $\boldsymbol{x}$, then*

$$\limsup_{r\to 0} \fint_{\substack{|\boldsymbol{y}-\boldsymbol{x}|\leq r,\\ (\boldsymbol{y}-\boldsymbol{x})\bullet\boldsymbol{\nu}>0}} |u(\boldsymbol{y}) - u^+(\boldsymbol{x})|^{\frac{d}{d-1}} = \limsup_{r\to 0} \fint_{\substack{|\boldsymbol{y}-\boldsymbol{x}|\leq r,\\ (\boldsymbol{y}-\boldsymbol{x})\bullet\boldsymbol{\nu}<0}} |u(\boldsymbol{y}) - u^-(\boldsymbol{x})|^{\frac{d}{d-1}} = 0.$$

- *If $\Omega$ has a smooth boundary, then $u$ satisfies general Poincaré inequalities. In particular, if either $\int_\Omega u = 0$ or $\mathrm{Trace}_{\partial\Omega}(u) = 0$, then*

$$\|u\|_1 \lesssim_\Omega \mathrm{TV}(u).$$

In other words, $u \in \mathbb{BV} \subset L^{\frac{d}{d-1}}$ is continuous almost everywhere up to a dimension $d-1$ set of discontinuities where the left and right trace are well defined almost everywhere.

### 1.3.4 Optimisation

As motivated at the start of Section 1.3, the search for an optimal reconstruction in inverse problems naturally links to the field of optimisation (Engl et al., 1996; Natterer, 2001). A standard example of this is (1.2).

In some relatively rare cases, minimisers can be computed directly. This is either the case for classical methods, or very specialised to a particular application. For instance, Tikhonov regularisation falls into this category and minimisers can be computed with standardised algorithms which require little user input (e.g. the MATLAB backslash operator). When direct methods are not available to solve (1.2), or not efficient enough, we turn to iterative methods where the fundamental goal is to produce a sequence $u_i \in \mathbb{U}$ such that $\mathrm{E}(u_i) \to \min_{u\in\mathbb{U}} \mathrm{E}(u)$ as fast as possible. The rest of this section is dedicated to introducing and motivating the core iterative methods which are currently used in inverse problems.

### Classical smooth methods

Most optimisation methods are based around the ideas of gradient descent. For this to be stated, we need a definition of smoothness.

**Definition 1.3.22.** *A function* $\mathrm{E}\colon \mathbb{U} \to \mathbb{R}$ *is called* $L$-smooth *if it is differentiable almost everywhere and*

$$\|\nabla \mathrm{E}(u) - \nabla \mathrm{E}(v)\|_2 \leq L \|u - v\|_2$$

*for all* $u, v \in \mathbb{U}$*.*

**Theorem 1.3.23.** *Suppose* $\mathrm{E}\colon \mathbb{U} \to \mathbb{R}$ *is* $L$*-smooth and* $u_0 \in \mathbb{U}$*. Gradient descent and Newton descent (when* $\mathrm{E} \in C^2(\mathbb{U})$*) are given by*

$$\begin{aligned} u_{n+1}^{GD} &:= u_n^{GD} - \frac{1}{L}\nabla \mathrm{E}(u_n^{GD}) & u_0^{GD} &= u_0, \\ u_{n+1}^{N} &:= u_n^{N} - \nabla^2 \mathrm{E}(u_n^{N})^{-1}\nabla \mathrm{E}(u_n^{N}) & u_0^{N} &= u_0, \end{aligned}$$

*respectively. For any* $u^*$ *with* $\nabla \mathrm{E}(u^*) = 0$*:*

- $$\mathrm{E}(u_n^{GD}) - \mathrm{E}(u^*) \lesssim \frac{L}{n} \|u_0 - u^*\|_2^2 .$$
- *if* $\mathrm{E}$ *is* $\mu$*-strongly convex, then*
  $$\mathrm{E}(u_n^{GD}) - \mathrm{E}(u^*) \lesssim (1 - \tfrac{\mu}{L})^n \|u_0 - u^*\|_2 .$$
- *if* $\mathrm{E} \in C^2$ *is* $\mu$*-strongly convex and* $\nabla^2 \mathrm{E}$ *is* $M$*-Lipschitz, then*
  $$\mathrm{E}(u_n^{N}) - \mathrm{E}(u^*) \lesssim \frac{\mu^3}{M^2} 2^{-2^{n-n_0+1}}$$
  *for some* $n_0 \in \mathbb{N}$*.*

These results demonstrate the spectrum of typical convergence speeds that one sees in optimisation. It is easy to design a general method which converges at a rate $\frac{1}{n}$, if there is strong convexity then linear rates are achieved, and methods capable of using second order derivatives can achieve super-linear rates although the analysis is much more challenging. The only convergence rate which can be improved (ignoring constants) is the rate $\frac{1}{n}$. The work of Nesterov gives explicit upper and lower rates which can be expected for convex functions.

**Theorem 1.3.24** (Nesterov (2004)). *Suppose* $\mathrm{E}\colon \mathbb{U}\to\overline{\mathbb{R}}$ *is convex and* $(u_n)_{n=1}^N$ *is a sequence in* $\mathbb{U}$ *for* $N\leq \frac{\dim(\mathbb{U})-1}{2}$.

- *If* $u_{n+1}\in u_0+\operatorname{span}\{\partial\mathrm{E}(u_0),\ldots,\partial\mathrm{E}(u_n)\}$*, then there exists constant* $C>0$ *and convex function* $\mathrm{E}'\colon\mathbb{U}\to\overline{\mathbb{R}}$ *such that*

$$\partial\mathrm{E}(u_n)=\partial\mathrm{E}'(u_n) \text{ for all } n\leq N, \qquad \mathrm{E}'(u_n)-\min_{u\in\mathbb{U}}\mathrm{E}'(u)\geq\frac{C}{\sqrt{n}}.$$

- *If* $\mathrm{E}\in C^1$ *and* $u_{n+1}\in u_0+\operatorname{span}\{\nabla\mathrm{E}(u_0),\ldots,\nabla\mathrm{E}(u_n)\}$*, then there exists constant* $C>0$ *and smooth convex function* $\mathrm{E}'\colon\mathbb{U}\to\overline{\mathbb{R}}$ *such that*

$$\nabla\mathrm{E}(u_n)=\nabla\mathrm{E}'(u_n) \text{ for all } n\leq N, \qquad \mathrm{E}'(u_n)-\min_{u\in\mathbb{U}}\mathrm{E}'(u)\geq\frac{C}{n^2}.$$

- *If* $\mathrm{E}$ *is* $L$*-smooth and* $\mu$*-strongly convex and* $u_{n+1}\in u_0+\operatorname{span}\{\nabla\mathrm{E}(u_0),\ldots,\nabla\mathrm{E}(u_n)\}$*, then there exists constant* $C>0$ *and* $L$*-smooth,* $\mu$*-strongly convex function* $\mathrm{E}'\colon\mathbb{U}\to\overline{\mathbb{R}}$ *such that*

$$\nabla\mathrm{E}(u_n)=\nabla\mathrm{E}'(u_n) \text{ for all } n\leq N, \qquad \mathrm{E}'(u_n)-\min_{u\in\mathbb{U}}\mathrm{E}'(u)\geq C\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^{2n}.$$

These results are sharp and achievable with simple accelerations of standard gradient descent.

**Theorem 1.3.25** (Nesterov (2004)). *Suppose* $\mathrm{E}\colon\mathbb{U}\to\mathbb{R}$ *and* $u_0\in\mathbb{U}$ *and fix sequence* $(\beta_n)_{n=1}^\infty$. *Subgradient descent is defined by*

$$u_{n+1}^{SD}\coloneqq u_n^{SD}-\frac{1}{\sqrt{n+1}}\frac{\varphi}{|\varphi|} \qquad \text{for some} \qquad \varphi\in\partial\mathrm{E}(u_n^{SD}), \qquad u_0^{SD}=u_0.$$

*If* $\mathrm{E}$ *is* $L$*-smooth, then Nesterov gradient descent is defined by*

$$u_{n+1}^{GD}\coloneqq v_n^{GD}-\frac{1}{L}\nabla\mathrm{E}(v_n^{GD}) \qquad \text{for} \qquad v_n^{GD}\coloneqq u_n^{GD}+\beta_n(u_n^{GD}-u_{n-1}^{GD}) \qquad u_0^{GD}=u_0.$$

*For any* $u^*$ *with* $0\in\partial\mathrm{E}(u^*)$*:*

- *Subgradient descent converges at the optimal rate,*

$$\mathrm{E}(u_n^{SD})-\mathrm{E}(u^*)\lesssim\frac{L}{\sqrt{n}}.$$

- *If* $\beta_n = \frac{t_n(1-t_n)}{t_n^2+t_{n+1}}$ *for the sequence* $t_n$ *defined by* $t_0 = 1$ *and* $t_{n+1}^2 = (1-t_{n+1})t_n^2$*, then Nesterov gradient descent converges at the optimal rate*

$$\mathrm{E}(u_n^{GD}) - \mathrm{E}(u^*) \lesssim \frac{1}{n^2}.$$

- *If* E *is* $\mu$*-strongly convex and* $\beta_n = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$*, then Nesterov gradient descent converges at the near-optimal rate*

$$\mathrm{E}(u_n^{GD}) - \mathrm{E}(u^*) \lesssim \left(1 - \sqrt{\tfrac{\mu}{L}}\right)^n.$$

### Classical non-smooth methods

While Nesterov accelerated methods are very simple and give optimal rates for smooth functions, in many practical applications E will not be smooth and the rate for subgradient descent is too slow. We have seen that (explicit) gradient descent converges stably whenever the Lipschitz constant is bounded. As in the numerical analysis of ODEs, if the system is not smooth enough for an explicit method to remain stable, then one should try an implicit version. In the case of convex optimisation, this is referred to as the proximal map.

**Definition 1.3.26.** *For a convex function* $\mathrm{g}\colon \mathbb{U} \to \overline{\mathbb{R}}$ *we define the* proximal map *of* g *at point* $u$ *by*

$$\mathrm{prox}_{\mathrm{g}}(u) := \operatorname*{argmin}_{v\in\mathbb{U}} \tfrac{1}{2}\left\|v-u\right\|_2^2 + \mathrm{g}(v).$$

*Equivalently,* $\mathrm{prox}_{\mathrm{g}}(u)$ *is the unique point such that*

$$\mathrm{prox}_{\mathrm{g}}(u) \in u - \partial\,\mathrm{g}(\mathrm{prox}_{\mathrm{g}}(u)).$$

While it is possible to apply proximal descent in the same manner as standard gradient descent, we shall present a generalised convergence result for functions of the form in (1.2).

**Algorithm 1.1** General Forward-Backward splitting

```
Suppose $\mathrm{E}(u) = \mathrm{f}(u) + \mathrm{g}(u)$ is a convex function where $\mathrm{f}\colon \mathbb{U} \to \mathbb{R}$ is smooth and $\mathrm{g}\colon \mathbb{U} \to \overline{\mathbb{R}}$
is convex.
Fix $\gamma > 0$, $(t_n)_{n=1}^\infty \subset [0,\infty)$ and $u_0 \in \mathbb{U}$.
$v_0 \leftarrow u_0, n \leftarrow 1$
repeat
    $v_n \leftarrow u_n + \frac{t_n - 1}{t_{n+1}}(u_n - u_{n-1})$
    $u_{n+1} \leftarrow \operatorname*{argmin}_{u\in\mathbb{U}} \frac{1}{2}\left\|u - v_n + \gamma\nabla\mathrm{f}(v_n)\right\|^2 + \gamma\,\mathrm{g}(u) = \mathrm{prox}_{\gamma\,\mathrm{g}}\left(v_n - \gamma\nabla\mathrm{f}(v_n)\right)$
    $n \leftarrow n + 1$
until converged
```

Before we state the convergence properties of Algorithm 1.1, we will define a popular form of acceleration referred to as FISTA.

**Definition 1.3.27.** *We say $(t_n)_{n\in\mathbb{N}}$ is a* FISTA stepsize choice *if*

$$t_n \geq 1, \qquad cn^2 \leq t_{n+1}^2 - t_{n+1} \leq t_n^2$$

*for some fixed $c > 0$, all $n \in \mathbb{N}$.*

The most common choices for $t_n$ are given by Chambolle and Dossal (2015), in particular they recommend the choice of $t_n = \frac{n+a-1}{a}$ for some $a \geq 2$. Liang and Schönlieb (2018) suggest several alternatives.

**Theorem 1.3.28** (Beck and Teboulle (2009))**.** *Suppose $\mathrm{E}(u) = \mathrm{f}(u) + \mathrm{g}(u)$ is a convex function where $\mathrm{f}\colon \mathbb{U} \to \mathbb{R}$ is $L$-smooth and $\mathrm{g}\colon \mathbb{U} \to \overline{\mathbb{R}}$ is convex.*

- *If $0 < \gamma \leq \frac{2}{L}$ and $t_n = 0$, then Algorithm 1.1 converges with*

$$\mathrm{E}(u_n) - \min_{u\in\mathbb{U}} \mathrm{E}(u) \lesssim \frac{1}{n}.$$

- *If $\mathrm{E}$ is $\mu$-strongly convex, $0 < \gamma \leq \frac{2}{L+\mu}$ and $t_n = 0$, then Algorithm 1.1 converges with*

$$\mathrm{E}(u_n) - \min_{u\in\mathbb{U}} \mathrm{E}(u) \lesssim \left(1 - \frac{\gamma(L+\mu)}{2}\frac{4\,{}^{\mu}\!/\!_{L}}{(1+{}^{\mu}\!/\!_{L})^2}\right)^{\frac{n}{2}}.$$

- *If $0 < \gamma \leq \frac{1}{L}$ and $(t_n)_{n=1}^{\infty}$ is a FISTA stepsize choice, then Algorithm 1.1 converges with*

$$\mathrm{E}(u_n) - \min_{u\in\mathbb{U}} \mathrm{E}(u) \lesssim \frac{1}{n^2}.$$

In particular, FISTA provides an optimal convergence rate for all convex (smooth or non-smooth) functions, and the standard Forward-Backward splitting algorithm already achieves linear convergence for strongly convex functions. There are two extensions to this basic framework. Replacing the forward gradient step with another proximal step derives the Backward-Backward algorithm and its inertial form called Douglas-Rachford splitting (Combettes and Pesquet, 2011). Alternatively, there are many proposed accelerations to FISTA in the literature such as by Liang and Schönlieb (2018) who proposes the variants in Algorithms 1.2 and 1.3 which even allows the stepsize choice $t_n = \infty$.

### Primal-dual methods

The final class of popular algorithms for convex optimisation utilise the duality properties of convex functions. The core example is Algorithm 1.4 which generalises other primal-dual algorithms such as ADMM (Chambolle and Pock, 2011). The convergence results are as follows.

**Algorithm 1.2** Restarting FISTA (Liang and Schönlieb, 2018)

```
Suppose $\mathrm{E}(u) = \mathrm{f}(u)+\mathrm{g}(u)$ is a convex function where $\mathrm{f}\colon \mathbb{U} \to \mathbb{R}$ is $L$-smooth and $\mathrm{g}\colon \mathbb{U} \to \overline{\mathbb{R}}$ is convex.
Fix $\gamma = \frac{1}{L}$ and $u_0 \in \mathbb{U}$.
$v_0 \leftarrow u_0, n \leftarrow 1$
repeat
    $v_n \leftarrow u_n + \frac{t_n - 1}{t_{n+1}}(u_n - u_{n-1})$
    $u_{n+1} \leftarrow \mathrm{prox}_{\gamma\,\mathrm{g}}(v_n - \gamma\nabla\,\mathrm{f}(v_n))$
    if $\langle u_{n+1} - v_n,\ u_{n+1} - u_n\rangle \leq 0$ then
        $u_{n+1} \leftarrow u_n$                                  ▷ Restart
    end if
    $n \leftarrow n+1$
until converged
```

**Algorithm 1.3** Greedy FISTA (Liang and Schönlieb, 2018)

```
Suppose $\mathrm{E}(u) = \mathrm{f}(u)+\mathrm{g}(u)$ is a convex function where $\mathrm{f}\colon \mathbb{U} \to \mathbb{R}$ is $L$-smooth and $\mathrm{g}\colon \mathbb{U} \to \overline{\mathbb{R}}$ is convex.
Fix $\gamma \in \left[\frac{1}{L}, \frac{2}{L}\right)$, $\xi < 1$, $S \geq 1$ and $u_0 \in \mathbb{U}$.      ▷ Suggested $\gamma = \frac{1.3}{L}, \xi = 0.96, S = 1$
$v_0 \leftarrow u_0, n \leftarrow 1$
repeat
    $v_n \leftarrow u_n + 1(u_n - u_{n-1})$
    $u_{n+1} \leftarrow \mathrm{prox}_{\gamma\,\mathrm{g}}(v_n - \gamma\nabla\,\mathrm{f}(v_n))$
    if $\langle u_{n+1} - v_n,\ u_{n+1} - u_n\rangle \leq 0$ then
        $u_{n+1} \leftarrow u_n$                                  ▷ Restart
    else if $\|u_{n+1} - u_n\| \geq S\,\|u_1 - u_0\|$ then
        $\gamma \leftarrow \max(\xi\gamma, \frac{1}{L})$                   ▷ 'Safeguard'
    end if
    $n \leftarrow n+1$
until converged
```

**Theorem 1.3.29** (Chambolle and Pock (2011))**.** *Suppose* $\mathrm{E}(u) = \mathrm{f}(u) + \mathrm{g}(v) + \mathrm{h}^*(\mathcal{A}v)$ *where* $\mathcal{A}\colon \mathbb{U} \to \Phi^*$ *is linear,* $\mathrm{f}\colon \mathbb{U} \to \mathbb{R}$ *is* $L$*-smooth, and* $\mathrm{g}\colon \mathbb{U} \to \overline{\mathbb{R}}$ *and* $\mathrm{h}\colon \Phi \to \overline{\mathbb{R}}$ *are convex. Define the saddle function* $\mathrm{S}(u, \varphi) = \mathrm{f}(u) + \mathrm{g}(u) + \langle \mathcal{A}u,\ \varphi\rangle - \mathrm{h}(\varphi)$ *for all* $u \in \mathbb{U}$, $\varphi \in \Phi$.

- *If* $\frac{1}{\sigma}\left(\frac{1}{\tau} - L\right) \geq \|\mathcal{A}\|^2$ *and* $t_n = 1$*, then Algorithm 1.4 converges with*

$$\mathrm{E}(u_n) - \min_{u\in\mathbb{U}} \mathrm{E}(u) \lesssim \frac{1}{n}.$$

- *If* $\mathrm{g}$ *is* $\mu$*-strongly convex and*

$$t_{n+1} = \frac{1}{\sqrt{1+\mu\tau_n}}, \qquad \tau_{n+1} = t_{n+1}\tau_n, \qquad \sigma_{n+1} = \frac{\sigma_n}{t_{n+1}}$$

*for* $\tau_0 = \frac{1}{2L}$*,* $\sigma_0 = \frac{L}{\|\mathcal{A}\|^2}$ *(or* $\tau_0 = \sigma_0 = \frac{1}{\|\mathcal{A}\|}$ *if* $L = 0$*), then Algorithm 1.4 converges with*

$$\mathrm{E}(u_n) - \min_{u\in\mathbb{U}} \mathrm{E}(u) \lesssim \frac{1}{n^2}.$$

- *If* $\mathrm{g}$ *is* $\mu_{\mathrm{g}}$*-strongly convex,* $\mathrm{h}$ *is* $\mu_{\mathrm{h}}$*-strongly convex and*

$$\frac{1}{t} = 1 + \mu_{\mathrm{g}}\tau = 1 + \mu_{\mathrm{h}}\sigma, \qquad t\sigma\, \|\mathcal{A}\|^2 \leq \frac{1 - L\tau}{\tau}$$

*then Algorithm 1.4 converges with*

$$\mathrm{E}(u_n) - \min_{u\in\mathbb{U}} \mathrm{E}(u) \lesssim t^n.$$

**Algorithm 1.4** Primal-dual iteration (a.k.a Chambolle-Pock algorithm) (Chambolle and Pock, 2011)

```
Suppose $\mathrm{S}(u,\varphi) = \mathrm{f}(u) + \mathrm{g}(u) + \langle \mathcal{A}u,\ \varphi\rangle - \mathrm{h}(\varphi)$ where $\mathcal{A}\colon \mathbb{U} \to \Phi^*$ is linear, $\mathrm{f}\colon \mathbb{U} \to \mathbb{R}$ is
$L$-smooth, and $\mathrm{g}\colon \mathbb{U} \to \overline{\mathbb{R}}$ and $\mathrm{h}\colon \Phi \to \overline{\mathbb{R}}$ are convex.
Fix sequences $\sigma_n, \tau_n, t_n > 0$, $u_0 \in \mathbb{U}$ and $\varphi \in \Phi$.
$\overline{u}_0 \leftarrow u_0,\ n \leftarrow 0$
repeat
    $u_{n+1} \leftarrow \operatorname*{argmin}_{u\in\mathbb{U}} \langle \nabla \mathrm{f}(\overline{u}_n),\ u - \overline{u}_n\rangle + \mathrm{g}(u) + \langle u,\ \mathcal{A}^*\varphi_n\rangle + \frac{1}{2\tau}\|u - \overline{u}_n\|_2^2$
                                    $= \mathrm{prox}_{\tau\,\mathrm{g}}(\overline{u}_n - \tau\nabla \mathrm{f}(\overline{u}_n) - \tau\mathcal{A}^*\varphi_n)$
    $\overline{u}_{n+1} \leftarrow u_{n+1} + t_n(u_{n+1} - u_n)$
    $\varphi_{n+1} \leftarrow \operatorname*{argmin}_{\varphi\in\Phi} \mathrm{h}(\varphi) - \langle \mathcal{A}u_{n+1},\ \varphi\rangle + \frac{1}{2\sigma}\|\varphi - \varphi_n\|_2^2$
                                    $= \mathrm{prox}_{\sigma\,\mathrm{h}}(\varphi_n + \sigma\mathcal{A}u_{n+1})$
    $n \leftarrow n + 1$
until converged
```

## 1.4 Electron microscopy preliminaries

Section 1.2.2 has motivated that there are many interactions occurring inside the electron microscope. For the purposes of solving inverse problems, we would like to have access to simple and efficient forward models which quantify the observed data. Everything *can* be simulated using methods such as Monte Carlo simulation (Demers et al., 2011), however, these work at a purely discrete level and so are extremely slow to run and give little insight for building simpler forward models.

The alternative is to start with the principle of least action and build a (continuum) Schrödinger equation. In this treatment, every atom and electron becomes a 'wave' of charge density and their electrostatic interactions are described by a PDE. There are two key challenges with this approach:

- Modelling each atom and electron independently is computationally slow and relatively uninformative. In practice atoms are assumed to be stationary.
- Even a full Schrödinger equation does not model inelastic events accurately. The creation and absorption of photons are discrete events which are governed by quantum field theory which is hard to merge with a Schrödinger equation.

The transition from atomic to continuum modelling, and the assumptions therein, is covered by density functional theory (often abbreviated to DFT) and provides the leading simulation methods for many areas of computational physics and chemistry (Giustino, 2014). Unfortunately, from the point of view of electron tomography, each electron probe requires its own simulation. In a typical application, this might require a non-trivial three dimensional PDE to be solved of order $10^6$ times for a single forward projection of a single sample. These methods are discussed in Section 1.4.1, however are prohibitively expensive for general inverse problems and so further simplifications are introduced in Sections 1.4.2, 1.4.4, 1.4.6 and 1.4.8 to improve computation time with minimal effect on accuracy. Section 1.4.3 contains discussion of popular computational methods arising from the content of Section 1.4.2. Section 1.4.7 introduces the concept of the limited angle problem which is a source of reconstruction artifacts in most electron tomography inverse problems.

The main aim of this section is to provide the details of the forward problems which will be used to formulate inverse problems in later chapters. On that basis, each model *can* be used to solve inverse problems but within the scope of this thesis we shall categorise each model as either fast or slow. Sections 1.4.6 to 1.4.8 discuss models which are used in this work to solve inverse problems, the earlier models in Sections 1.4.1 to 1.4.5 are more accurate but also much slower. In Chapter 3 we will use two of the slow methods from Section 1.4.3 to quantify the accuracy of a newly proposed forward model approximation. The remaining models are described here to provide background and context for general simulation models.

### 1.4.1 PDE models

There are three scales at which one can consider materials in an EM:

1. the continuum where individual atoms are too small to be seen,
2. the atomic where atoms are spherical blobs,
3. and the sub-atomic where atoms and bound electrons form more complex structures.

Density functional theory covers most of this range and provides effective techniques for studying a broad class of samples from nanostructures to molecules to solids. This topic is discussed at length in (Giustino, 2014, Chapter 2) and this section aims to provide a condensed summary.

At the heart of density functional theory are many-body Schrödinger equations to simulate (electrostatic) material properties. It is the idea that all materials are a sum of negatively charged electrons and positively charged nuclei which interact exclusively according to the electrostatic Coulomb forces. The three (scalar) potentials of interest, for two particles offset by a vector $\boldsymbol{r}$, are:

$$\begin{aligned}\text{Electron-Electron repulsion, } E_{ee}(\boldsymbol{r}) &= \frac{e^2}{4\pi\varepsilon_0}\times\frac{1}{|\boldsymbol{r}|}\\ \text{Nuclei-Nuclei repulsion, } E_{nn}(\boldsymbol{r}) &= \frac{e^2}{4\pi\varepsilon_0}\times\frac{Z_1Z_2}{|\boldsymbol{r}|}\\ \text{Electron-Nuclei attraction, } E_{en}(\boldsymbol{r}) &= \frac{e^2}{4\pi\varepsilon_0}\times -\frac{Z_1}{|\boldsymbol{r}|}\end{aligned}$$

where $Z_i$ are the number of protons in each nuclei, $e$ is the charge of one electron/proton and $\varepsilon_0$ is the permittivity of vacuum. If there are $N$ electrons at position $\boldsymbol{r}_i \in \mathbb{R}^3$ with mass $m_e$ and $M$ nuclei at positions $R_i \in \mathbb{R}^3$, atomic number $Z_i$ and mass $M_i$, then we get the total potential

$$V(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N,\boldsymbol{R}_1,\ldots,\boldsymbol{R}_M) = \frac{e^2}{4\pi\varepsilon_0}\left[\sum_{\substack{i,j\in[N]\\ i\neq j}}\frac{1}{|\boldsymbol{r}_i-\boldsymbol{r}_j|} - 2\sum_{\substack{i\in[N],\\ j\in[M]}}\frac{Z_j}{|\boldsymbol{r}_i-\boldsymbol{R}_j|} + \sum_{\substack{i,j\in[M],\\ i\neq j}}\frac{Z_iZ_j}{|\boldsymbol{R}_i-\boldsymbol{R}_j|}\right].$$

The total energy $T_E > 0$ is defined to be the sum of potential and kinetic energy, therefore the corresponding Schrödinger equation is

$$-\frac{\hbar^2}{2}\sum_{i\in[N]}\frac{\Delta_{\boldsymbol{r}_i}\Psi(\boldsymbol{r},\boldsymbol{R})}{m_e} - \frac{\hbar^2}{2}\sum_{i\in[M]}\frac{\Delta_{\boldsymbol{R}_i}\Psi(\boldsymbol{r},\boldsymbol{R})}{M_i} + V(\boldsymbol{r},\boldsymbol{R})\Psi(\boldsymbol{r},\boldsymbol{R}) = T_E\Psi(\boldsymbol{r},\boldsymbol{R})$$

for each $\boldsymbol{r} \in \mathbb{R}^{3N}$, $\boldsymbol{R} \in \mathbb{R}^{3M}$, and where $\hbar$ is Planck's constant. After a change of units to the *Hartree atomic units*, the equation simplifies to

$$\left[-\textstyle\sum_{i\in[N]} \frac{\Delta_{\boldsymbol{r}_i}}{2} - \sum_{i\in[M]} \frac{\Delta_{\boldsymbol{R}_i}}{2} - \sum_{\substack{i\in[N],\\ j\in[M]}} \frac{Z_j}{|\boldsymbol{r}_i-\boldsymbol{R}_j|} + \frac{1}{2}\sum_{\substack{i,j\in[N]\\ i\neq j}} \frac{1}{|\boldsymbol{r}_i-\boldsymbol{r}_j|} + \frac{1}{2}\sum_{\substack{i,j\in[M],\\ i\neq j}} \frac{Z_iZ_j}{|\boldsymbol{R}_i-\boldsymbol{R}_j|}\right]\Psi = T_E\Psi.$$

From a functional analysis point of view, the left-hand side is a symmetric, positive definite operator with minimal eigenvalue $T_E$ referred to as the *total energy* and $L^2$-normalised eigenvector $\Psi$ referred to as the *wavefunction*. The physical interpretation of $\Psi\colon \mathbb{R}^{3N+3M} \to \mathbb{C}$ is that $|\Psi|^2$ is the probability measure of finding all electrons at points $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N$ and each nuclei at $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_M$. It is the normalised eigenvector so $\int |\Psi|^2 = 1$ which confirms that this is the correct normalisation. More formally, this interpretation is referred to as *Born's rule* which is a fundamental axiom to quantum mechanics. To convert this into more physical quantities, for example, we can compute the average charge $\rho\colon \mathbb{R}^3 \to \mathbb{R}$ by

$$\rho(\boldsymbol{x}) = \sum_{i\in[N]} -e\int_{\boldsymbol{r}_i=\boldsymbol{x}} |\Psi(\boldsymbol{r},\boldsymbol{R})|^2 + \sum_{i\in[M]} Z_i e\int_{\boldsymbol{R}_i=\boldsymbol{x}} |\Psi(\boldsymbol{r},\boldsymbol{R})|^2.$$

This model is very accurate but prohibitively expensive when the number of atoms gets large. A $10\,\mathrm{nm}^3$ silicon crystal would require $N + M \sim 10^9$, and this is still too small for typical microscopy. To reduce complexity, the first assumption is the *clamped nuclei* assumption which asserts

$$\Psi(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_N) = \Psi(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)\delta_{\overline{\boldsymbol{R}}_1}(\boldsymbol{R}_1)\ldots\delta_{\overline{\boldsymbol{R}}_M}(\boldsymbol{R}_M)$$

for some 'clamps' $\overline{\boldsymbol{R}}_i \in \mathbb{R}^3$. Further modifications focus on the simplification of the electron-electron interactions for numerical efficiency whilst maintaining physical accuracy. This results in the Kohn-Sham equations (Kohn and Sham, 1965) which require computation of $N$ eigenfunctions of a three-dimensional PDE of the form $[-\Delta + \overline{V}(\boldsymbol{r})]\psi(\boldsymbol{r}) = T_E\psi(\boldsymbol{r})$ rather than a single eigenfunction of a $3N$-dimensional PDE. This is by no means the end of the story, but outlines the modern approach to, so called, *first principles* materials simulation.

### 1.4.2 Born approximations

Another approach in EM simulation is to consider atoms as fixed (time independent) and the interaction with electrons in the beam are explained as a scattering problem. This section aims to briefly summarise the necessary content from Cowley (1981) which discusses diffraction microscopy in much greater detail. Diffraction theory was first derived for light waves (e.g. X-rays) satisfying the Boltzmann equations, however, later derivations generalised this for both electron and neutron scattering, utilising the fact that all quantum particles act as waves. The methods of density functional theory are used to compute the charge density of the fixed sample, $\rho\colon \mathbb{R}^3 \to \mathbb{R}$, and this is converted to a scattering potential, $V\colon \mathbb{R}^3 \to \mathbb{R}$ by $\Delta V = -4\pi\rho$. In this scenario, the only degree of freedom is the electron wave $\psi\colon \mathbb{R}^3 \to \mathbb{C}$ generated by the microscope, the corresponding PDE is

$$\left[\Delta + \frac{4\pi^2}{\lambda^2} + \mu V(\boldsymbol{r})\right]\psi(\boldsymbol{r}) = 0 \tag{1.4}$$

where $\lambda$ is the electron wavelength and $\mu > 0$ is a coupling parameter. Suppose that the electrons are emitted from $(0,0,0)$ in a plane wave down the $z$-axis with wave vector $\boldsymbol{k}_0 = (0,0,\frac{4\pi}{\lambda})$ and the detector is contained in the plane at $z = Z$. The solution can be written implicitly in terms of the Green's function

$$\psi(\boldsymbol{r}) = \psi^{(0)}(\boldsymbol{r}) + \mu\int_{r'_z\in[0,Z]} \frac{\exp\left(\imath|\boldsymbol{k}_0||\boldsymbol{r}-\boldsymbol{r}'|\right)}{4\pi|\boldsymbol{r}-\boldsymbol{r}'|} V(\boldsymbol{r}')\psi(\boldsymbol{r}')d\boldsymbol{r}'.$$

The Born series expansion can be derived as a Picard iteration on this expression:

$$\begin{aligned}
\psi^{(0)}(\boldsymbol{r}) &= \exp\left(-\imath\boldsymbol{k}_0\cdot\boldsymbol{r}\right)\\
\sum_{i=0}^{n+1}\psi^{(i)}(\boldsymbol{r}) &= \psi^{(0)}(\boldsymbol{r}) + \frac{\mu}{4\pi}\sum_{i=0}^{n}\int_{r'_z\in[0,Z]} \frac{\exp\left(\imath|\boldsymbol{k}_0||\boldsymbol{r}-\boldsymbol{r}'|\right)}{|\boldsymbol{r}-\boldsymbol{r}'|} V(\boldsymbol{r}')\psi^{(i)}(\boldsymbol{r}')d\boldsymbol{r}'.
\end{aligned}$$

For weakly scattering samples (thin with small $\lambda$), this series is said to converge very quickly and only the first order Born approximation is used in practice. A further approximation is made based on the fact that size of samples are in the scale of 100 nm and the distance to the detector is greater than 10 cm. If $V$ is supported in the ball of radius $R$ for $R \ll Z$, then for $\boldsymbol{r} = (x,y,Z)$,

$$|\boldsymbol{r}-\boldsymbol{r}'| = |\boldsymbol{r}|\sqrt{\left|\frac{\boldsymbol{r}}{|\boldsymbol{r}|} - \frac{\boldsymbol{r}'}{|\boldsymbol{r}|}\right|^2} \sim |\boldsymbol{r}|\sqrt{1 - 2\frac{\boldsymbol{r}\cdot\boldsymbol{r}'}{|\boldsymbol{r}|^2}} \sim |\boldsymbol{r}| - \frac{\boldsymbol{r}\cdot\boldsymbol{r}'}{|\boldsymbol{r}|}$$

for all $\boldsymbol{r}' \in \operatorname{supp}(V)$. Substituting this into the formula for $\psi^{(1)}$ gives

$$\begin{aligned}
\psi^{(1)}(\boldsymbol{r}) &= \frac{\mu}{4\pi}\int_{r'_z\in[0,Z]} \frac{\exp\left(\imath|\boldsymbol{k}_0||\boldsymbol{r}-\boldsymbol{r}'|\right)}{|\boldsymbol{r}-\boldsymbol{r}'|} V(\boldsymbol{r}')\exp\left(-\imath\boldsymbol{k}_0\cdot\boldsymbol{r}'\right)d\boldsymbol{r}'\\
&\sim \frac{\mu}{4\pi}\frac{\exp\left(\imath|\boldsymbol{k}_0||\boldsymbol{r}|\right)}{|\boldsymbol{r}|}\int_{|\boldsymbol{r}'|\leq R} \frac{\exp\left(-\imath\frac{|\boldsymbol{k}_0|}{|\boldsymbol{r}|}\boldsymbol{r}\cdot\boldsymbol{r}'\right)}{1-\frac{\boldsymbol{r}\cdot\boldsymbol{r}'}{|\boldsymbol{r}|}} V(\boldsymbol{r}')\exp\left(-\imath\boldsymbol{k}_0\cdot\boldsymbol{r}'\right)d\boldsymbol{r}'\\
&= \frac{\mu}{4\pi}\frac{\exp\left(\imath|\boldsymbol{k}_0||\boldsymbol{r}|\right)}{|\boldsymbol{r}|}\int_{|\boldsymbol{r}'|\leq R} \frac{1}{1-\frac{\boldsymbol{r}\cdot\boldsymbol{r}'}{|\boldsymbol{r}|}} V(\boldsymbol{r}')\exp\left(-\imath\left[|\boldsymbol{k}_0|\frac{\boldsymbol{r}}{|\boldsymbol{r}|} + \boldsymbol{k}_0\right]\cdot\boldsymbol{r}'\right)d\boldsymbol{r}'\\
&\approx \frac{\mu}{4\pi}\frac{\exp\left(-\imath\boldsymbol{k}_0\cdot\boldsymbol{r}\right)}{|\boldsymbol{r}|}\mathcal{F}[V]\left(|\boldsymbol{k}_0|\frac{\boldsymbol{r}}{|\boldsymbol{r}|} + \boldsymbol{k}_0\right).
\end{aligned}$$

The last line shows that diffraction 'looks like' a Fourier transform up to a first order correction and shifting of basis. This is a true first order expansion if $|x|, |y| \ll Z$ as well (i.e. *small angle scattering*).

Historically, this formula has been derived several times. One classical derivation names it the *Ewald sphere* model which highlights the important geometrical feature that

$$\left\{|\boldsymbol{k}_0|\frac{\boldsymbol{r}}{|\boldsymbol{r}|}+\boldsymbol{k}_0 \quad \text{s.t.} \quad \boldsymbol{r}\in\mathbb{R}^3\right\} \quad = \quad \left\{\boldsymbol{k}\in\mathbb{R}^3 \quad \text{s.t.} \quad |\boldsymbol{k}-\boldsymbol{k}_0|\leq|\boldsymbol{k}_0|=\frac{2\pi}{\lambda}\right\}.$$

This shows that diffraction imaging is like sampling the Fourier transform on a sphere (called *the Ewald sphere*). We now need to convert this to an observed image, a diffraction pattern $D\colon\mathbb{R}^2\to[0,\infty)$. This is done in three stages

1. By Born's rule, the observed number of electrons at point $(x,y,-Z)$ is proportional to $\left|\psi\left(|\boldsymbol{k}_0|\frac{(x,y,-Z)}{|(x,y,-Z)|}+\boldsymbol{k}_0\right)\right|^2$

2. If $Z$ is unknown, then we suppose the coordinates on the detector are already in *reciprocal space* (a.k.a Fourier space) and project pixel $\boldsymbol{k}=(k_x,k_y)$ to point $\boldsymbol{K}\in\mathbb{R}^3$ on the sphere. This samples the Fourier transform at the point

$$\boldsymbol{K}=\boldsymbol{k}_0+|\boldsymbol{k}_0|\frac{(k_x,k_y,0)-\boldsymbol{k}_0}{|(k_x,k_y,0)-\boldsymbol{k}_0|}=\frac{\left(k_x,k_y,\left[\sqrt{1+\frac{\lambda^2}{4\pi^2}|\boldsymbol{k}|^2}-1\right]\frac{2\pi}{\lambda}\right)}{\sqrt{1+\frac{\lambda^2}{4\pi^2}|\boldsymbol{k}|^2}}.$$

   Up to first order terms for $|\boldsymbol{k}|\ll\frac{2\pi}{\lambda}$, this is equivalent to

$$\boldsymbol{K}=\left(k_x,k_y,\left[1-\sqrt{1-\frac{\lambda^2}{4\pi^2}|\boldsymbol{k}|^2}\right]\frac{2\pi}{\lambda}\right)=:(\boldsymbol{k},k_z(\boldsymbol{k})).$$

3. If the electron beam $\psi^{(0)}(\boldsymbol{r})=\Psi_p(\boldsymbol{r})\exp(-\imath\boldsymbol{k}_0\cdot\boldsymbol{r})$ where $\Psi_p\colon\mathbb{R}^2\to\mathbb{C}$ (called the *probe function*) is constant in the $z$-direction, then the first order perturbation becomes

$$\psi^{(1)}(\boldsymbol{r})\sim\frac{\mu}{4\pi}\frac{\exp(\imath|\boldsymbol{k}_0||\boldsymbol{r}|)}{|\boldsymbol{r}|}\mathcal{F}[\Psi_pV]\left(|\boldsymbol{k}_0|\frac{\boldsymbol{r}}{|\boldsymbol{r}|}+\boldsymbol{k}_0\right).$$

Combining these gives the formal Ewald sphere model for diffraction imaging:

$$D(\boldsymbol{k})=|\mathcal{F}[\Psi_pV](\boldsymbol{k},k_z(\boldsymbol{k}))|^2, \qquad k_z(\boldsymbol{k}):=\left[1-\sqrt{1-\frac{\lambda^2}{4\pi^2}|\boldsymbol{k}|^2}\right]\frac{2\pi}{\lambda}. \tag{1.5}$$

Note that there are a couple of ambiguities in this model. Firstly, there are two conventions of projection from 2D to 3D, as described above. More subtly, this is actually a first order perturbation rather than a first order approximation. Born's series suggests $\psi\approx\psi^{(0)}+\psi^{(1)}$ but we have dropped the zeroth-order term. If $\Psi_p$ is either constant (plane wave) or locally supported in $x$ and $y$, then the difference is negligible. In the former case, there is a constant offset and in the latter, the effect is only noticed in the direct beam at $\boldsymbol{k}\approx 0$. Physically, thicker samples diffract more so the direct beam is dimmer, as predicted by the first order expansion. On the other hand, in (1.5) the direct beam looks brighter. In either case, this is a subtle effect that can typically be ignored.

A final common modification of this approximation is achieved as a high-energy limit of (1.5). If $\lambda \approx 0$, then the Ewald sphere is very flat. Asymptotically, the sphere becomes a hyperplane:

$$k_z(\boldsymbol{k}) = \left[1 - \sqrt{1 - \frac{\lambda^2}{4\pi^2}|\boldsymbol{k}|^2}\right]\frac{2\pi}{\lambda} \sim \frac{\lambda}{4\pi}|\boldsymbol{k}|^2 \xrightarrow{\lambda\to 0} 0.$$

In this limit, diffraction becomes a 2D problem:

$$\lim_{\lambda\to 0} D(\boldsymbol{k}) = |\mathcal{F}[\Psi_p V]\,(\boldsymbol{k},0)|^2 = \left|\int_{\mathbb{R}^2} \Psi_p(x,y)\left[\int_{\mathbb{R}} V(x,y,z)dz\right]\exp\left[-\imath\boldsymbol{k}\cdot(x,y)\right]dxdy\right|^2. \quad (1.6)$$

This shows that diffraction is less sensitive to variation in the $z$ direction, especially if the wavelength and thickness of the sample are small ($\lambda$ and $R$). For high-energy (ca. 60 keV to 300 keV) incident electrons, the corresponding de Broglie wavelength is ca. 0.0487 Å to 0.0197 Å and so the approximation is reasonably accurate.

### 1.4.3 Simulation

There are three core simulation methods for electron scattering derived from the previously described models. We provide a brief introduction and comparison of these simulation methods here and further details can be found in Kirkland (1998).

#### Ewald sphere

The Ewald sphere model has already been introduced in (1.5) which should be thought of as the exact solution of the linearisation of the PDE in (1.4). The simplified version in (1.6) describes solves the solution to the linearised problem in the limit where the electron is very fast.

#### Bloch waves

Bloch waves are a functional analytic approach to solve the Schrödinger equation (1.4) assuming that the sample is fixed on a periodic domain. If there is only a single electron in the electron beam and the atomic potential is also periodic, then the eigenfunctions of (1.4) can typically be written down analytically. The generic electron beam is then decomposed into this basis of eigenfunctions leading to the analytical form of the diffraction pattern.

### Multislice

Multislice is another technique which starts with (1.4) and then assumes the electrons are sufficiently fast that the PDE becomes an ODE in $z$ of the form

$$\nabla_z \psi = [\mathcal{A} + B]\,\psi$$

where $\mathcal{A} \propto \imath\Delta_{x,y}$ and $B \propto \imath\left(V + \frac{4\pi^2}{\lambda^2}\right)$. This ODE is still not exactly solvable but much easier to discretise. The volume is divided into thin 'slices' and on each slice the propagation

$$\psi(x,y,z+\Delta z) = \exp(\Delta z\mathcal{A})\exp\left(\int_z^{z+\Delta z} B(x,y,\widetilde{z})d\widetilde{z}\right)\psi(x,y,z)$$

is computed. While this is a relatively complex formula, it is also a relatively standard operator splitting method applied to the ODE. More advanced solvers have also been proposed and are now in commonly used software packages such as by Lobato and Van Dyck (2015).

### Comparison

The Ewald sphere model, Bloch waves, and multislice are all common methods for simulation of diffraction patterns as each has its own advantages.

Bloch waves give the most accurate analytical information but can only be computed exactly under a lot of extra assumptions. In particular, the atomic potential is ideally exactly periodic which excludes all non-trivial samples. Bloch waves can be computed numerically for generic potentials but then the analytical benefits are lost.

The Ewald sphere model provides the best geometrical intuition, linking diffraction directly to the Fourier transform. The biggest limitation is that each electron is allowed to scatter at most once, placing it in the category of *kinematical* models. The other models allow multiple scattering events and are therefore *dynamical* models. The other limitation is that the full Ewald sphere model has a relatively large computational complexity, dominated by the computation of a large 3D Fourier transform. The high energy limit form of the Ewald sphere model is much faster, requiring a single 2D Fourier transform, but captures fewer of the geometrical features of diffraction.

Multislice is the most popular method for experimental diffraction simulation because it is relatively efficient, comparable to the full Ewald sphere, and is quantitatively accurate.

The biggest limitation of all methods considered here is in the accounting of inelastic scattering. This is a fundamental limitation of the Schrödinger equation (1.4) because energy of the electron must be conserved. In reality, other quantum processes occur to draw energy out of the system. The main result of this is that the spectral microscopy cannot be simulated from this equation and there is a quantitative error in all diffraction intensities.

### 1.4.4 Diffraction of crystals

Up to this point we have discussed simulation of diffraction patterns from generic samples. Crystals are a common subclass of possible specimens which posses highly structured diffraction patterns. In particular, a crystal is defined as a material that has an 'essentially sharp' diffraction pattern (Committee, 1992). This means that diffraction images look like a sparse sum of sharp *Bragg peaks* or *Bragg discs*, although in practice there will always be background intensity between the peaks. Through (1.5) we can infer that sparse diffraction patterns correspond to sparse Fourier transforms. This leads to the definition of an ideal crystal.

**Definition 1.4.1.** *The electrostatic potential of an* ideal crystal $u_0\colon \mathbb{R}^3 \to \mathbb{R}$ *satisfies*

$$\mathcal{F}[u_0] = \sum_{i=1}^{\infty} a_i \delta_{\boldsymbol{p}_i}$$

*for some weightings* $a_i \in \mathbb{C}$ *and* Bragg peaks $\boldsymbol{p}_i \in \mathbb{R}^3$*.*

*If*

$$\{\boldsymbol{p}_i \text{ s.t. } i \in \mathbb{N}\} = \left\{B\boldsymbol{j} \text{ s.t. } \boldsymbol{j} \in \mathbb{Z}^3\right\}$$

*for some* $B \in \mathbb{R}^{3\times 3}$*, then* $u_0$ *is called a* conventional crystal*. The columns of* $B$ *are called the* reciprocal lattice vectors *of* $u_0$*.*

Sparse Fourier transforms correspond to periodic potentials, as is made precise by the following lemma.

**Lemma 1.4.2.** *If* $u_0$ *is a conventional crystal with reciprocal lattice vectors* $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3$*, then*

$$u_0\left(\boldsymbol{x} + \frac{2\pi}{\det(B)}\boldsymbol{b}_i \times \boldsymbol{b}_j\right) = u_0(\boldsymbol{x})$$

*for all* $\boldsymbol{x} \in \mathbb{R}^3$ *and* $i, j \in [3]$*.*

*Proof.* By direct computation,

$$u_0(\boldsymbol{x}) = (2\pi)^{-3} \sum_{i\in\mathbb{N}} a_i \int_{\mathbb{R}^3} \delta_{\boldsymbol{p}_i}(\boldsymbol{k}) e^{\imath \boldsymbol{x}\cdot\boldsymbol{k}} = (2\pi)^{-3} \sum_{i\in\mathbb{N}} a_i e^{\imath\boldsymbol{x}\cdot\boldsymbol{p}_i}.$$

For each $\boldsymbol{p}_i = \sum_{j=1}^3 i_j \boldsymbol{b}_j$, then

$$\exp\left(\imath \frac{2\pi}{\det(B)}(\boldsymbol{b}_1 \times \boldsymbol{b}_2)\cdot\boldsymbol{p}_i\right) = \exp\left(\imath \frac{2\pi}{\det(B)}(\boldsymbol{b}_1 \times \boldsymbol{b}_2)\cdot(i_1\boldsymbol{b}_3)\right) = 1.$$

Therefore $u_0$ is a sum of $\frac{2\pi}{\det(B)}\boldsymbol{b}_1 \times \boldsymbol{b}_2$-periodic functions. The other cases follow by symmetry. □

The key point of this lemma is that the real domain and Fourier (or 'reciprocal') domain representations are equivalent, but each have distinct advantages. The real-space representation is convenient for discretisation into simple repeating blocks and the Fourier representation is optimal for computing kinematical diffraction patterns.

One example of this is the computation of the diffraction pattern of a truncated conventional crystal.

**Lemma 1.4.3.** *Define*

$$u(\boldsymbol{x}) := \mathbb{1}_{|\boldsymbol{x}|_\infty \leq \nicefrac{\rho}{2}} u_0(\boldsymbol{x}) = \begin{cases} u_0(\boldsymbol{x}) & \boldsymbol{x} \in [-\nicefrac{\rho}{2}, \nicefrac{\rho}{2}]^3 \\ 0 & \textit{else} \end{cases}$$

*where $\rho > 0$ is the width/depth of $u$ and let $D$ denote the diffraction pattern defined by the Ewald sphere model in (1.5). In this case, $D$ simplifies to*

$$D(\boldsymbol{k}) = \left| \sum_{i=1}^{\infty} a_i \rho \operatorname{sinc}(\rho[k_z(\boldsymbol{k}) - p_{i,z}]) f(k_x - p_{i,x}, k_y - p_{i,y}) \right|^2$$

*where*

$$f(\boldsymbol{k}) := \rho^2 \mathcal{F}[\Psi_p] \star [\operatorname{sinc}(\rho \cdot)](\boldsymbol{k}) = \rho^2 \int_{\mathbb{R}^2} \mathcal{F}[\Psi_p](\boldsymbol{k} - \boldsymbol{k}') \operatorname{sinc}(\rho \boldsymbol{k}') d\boldsymbol{k}'. \tag{1.7}$$

*Proof.* Recall the definition of $D$ from (1.5):

$$D(\boldsymbol{k}) = |\mathcal{F}[\Psi_p u](\boldsymbol{k}, k_z(\boldsymbol{k}))|^2 = |\mathcal{F}[\Psi_p \mathbb{1}_{|\boldsymbol{x}|_\infty \leq \nicefrac{\rho}{2}} u_0](\boldsymbol{k}, k_z(\boldsymbol{k}))|^2.$$

By the Fourier convolution theorem we can expand

$$\mathcal{F}\left[\Psi_p \mathbb{1}_{|\boldsymbol{x}|_\infty \leq \nicefrac{\rho}{2}}\right](\boldsymbol{K}) = \mathcal{F}[\Psi_p] \star \mathcal{F}\left[\mathbb{1}_{|\boldsymbol{x}|_\infty \leq \nicefrac{\rho}{2}}\right](\boldsymbol{K}) = \mathcal{F}[\Psi_p] \star \left[\rho^3 \operatorname{sinc}(\rho k)\right](\boldsymbol{K}).$$

As $\Psi_p(\boldsymbol{r}) = \Psi_p(r_x, r_y)$ only varies in the first two variables, this convolution splits over $\mathbb{R}^{2+1}$ to

$$\mathcal{F}\left[\Psi_p \mathbb{1}_{|\boldsymbol{x}|_\infty \leq \nicefrac{\rho}{2}}\right](\boldsymbol{K}) = \underbrace{\mathcal{F}[\Psi_p] \star \left[\rho^2 \operatorname{sinc}(\rho k_x) \operatorname{sinc}(\rho k_y)\right](K_x, K_y)}_{= f(K_x, K_y)} \rho \operatorname{sinc}(\rho K_z).$$

Recall $\mathcal{F}[u_0] = \sum_{i \in \mathbb{N}} a_i \delta_{\boldsymbol{p}_i}$, therefore the final convolution is just translation by $\boldsymbol{p}_i$. The result of this is

$$\begin{aligned} \mathcal{F}[\Psi_p u](\boldsymbol{K}) &= \sum_{i \in \mathbb{N}} a_i [\rho \operatorname{sinc}(\rho k_z) f(k_x, k_y)] \star \delta_{\boldsymbol{p}_i}(\boldsymbol{K}) \\ &= \sum_{i \in \mathbb{N}} a_i \rho \operatorname{sinc}(\rho(K_z - p_{i,z})) f(K_x - p_{i,x}, K_y - p_{i,y}) \end{aligned}$$

as required. □

Lemma 1.4.3 clarifies what is meant by the idea of ‘essentially sharp’ diffraction patterns. The diffraction pattern $D$ is a sum of spikes of shape $f$ centred at $(p_{i,x}, p_{i,y})$ and of intensity

$a_i\rho\,\mathrm{sinc}(k_z(\boldsymbol{k})-\rho p_{i,z})$. In the high energy limit (1.6), $k_z=0$ and this is just a constant damping factor of $\mathrm{sinc}(\rho p_{i,z})\leq 1$ which suppresses the intensity of the signal coming from spikes such that $p_{i,z}\neq 0$.

### 1.4.5 Precession diffraction imaging

Precession, or more precisely *double-conical electron beam rocking*, is a microscopy technique used to 'simplify' the forward model of diffraction proposed by Vincent and Midgley (1994). The procedure is formally referred to as *precession electron diffraction* (PED). The schematic is given in Figure 1.8 which shows the electron beam tilted away from the optical axis ($z$-axis) by a precession angle $\alpha$ then rotated above and below the sample. Heuristically, it is hoped that this blurring averages out the higher order Born terms resulting in 'more kinematical' diffraction patterns. Figure 1.9 shows examples of diffraction patterns obtained with kinematical/dynamical simulators with and without precession. The take-home message from this figure is that 'exact' diffraction is messy and complex but exact diffraction with precession is very close to kinematical diffraction with precession. This is a great aid in building practical models for diffraction.

Analytically, it is simpler to consider the sample being precessed and everything else remaining stationary. This is essentially the role of the second beam deflection, to mimick the detector precessing in anti-phase to the beam. With this interpretation, we can express the intensities of a PED pattern, $D_\alpha(\boldsymbol{k})$, as,

$$\begin{aligned} D_\alpha(\boldsymbol{k}) = \mathop{\mathbb{E}}_{t}\Bigg\{ &\left|\mathcal{F}[\Psi_p u(R_t\boldsymbol{x})]\right|^2(\boldsymbol{k},k_z(\boldsymbol{k})) \text{ such that } t\in[0,2\pi) \text{ and} \\ & R_t = \begin{pmatrix}\cos(t) & \sin(t) & 0\\ -\sin(t) & \cos(t) & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\ 0 & \cos(\alpha) & \sin(\alpha)\\ 0 & -\sin(\alpha) & \cos(\alpha)\end{pmatrix}\begin{pmatrix}\cos(t) & -\sin(t) & 0\\ \sin(t) & \cos(t) & 0\\ 0 & 0 & 1\end{pmatrix}\Bigg\}. \end{aligned} \tag{1.8}$$

There is no intrinsic randomness but considering $t\sim\mathrm{Uniform}[0,2\pi)$ ensures that the expectation is the desired value.

Heuristically, standard diffraction patterns have complex features which are hard to model, for instance the non-uniform intensities in Figure 1.9a, so we introduce a small blurring effect. This must be large enough to smooth, but not too large to invalidate (1.5). The typical range considered in the literature is $\alpha\in[0.5°,2°]$.

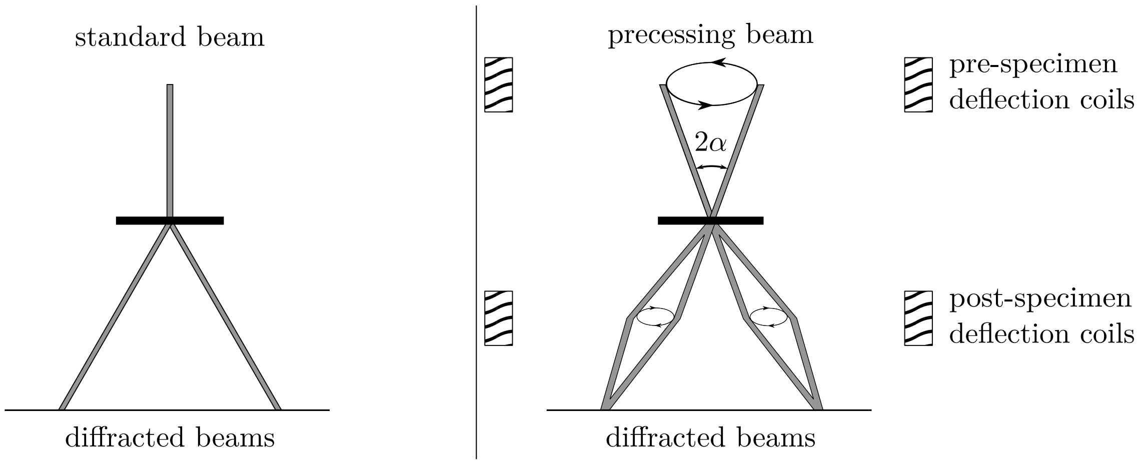
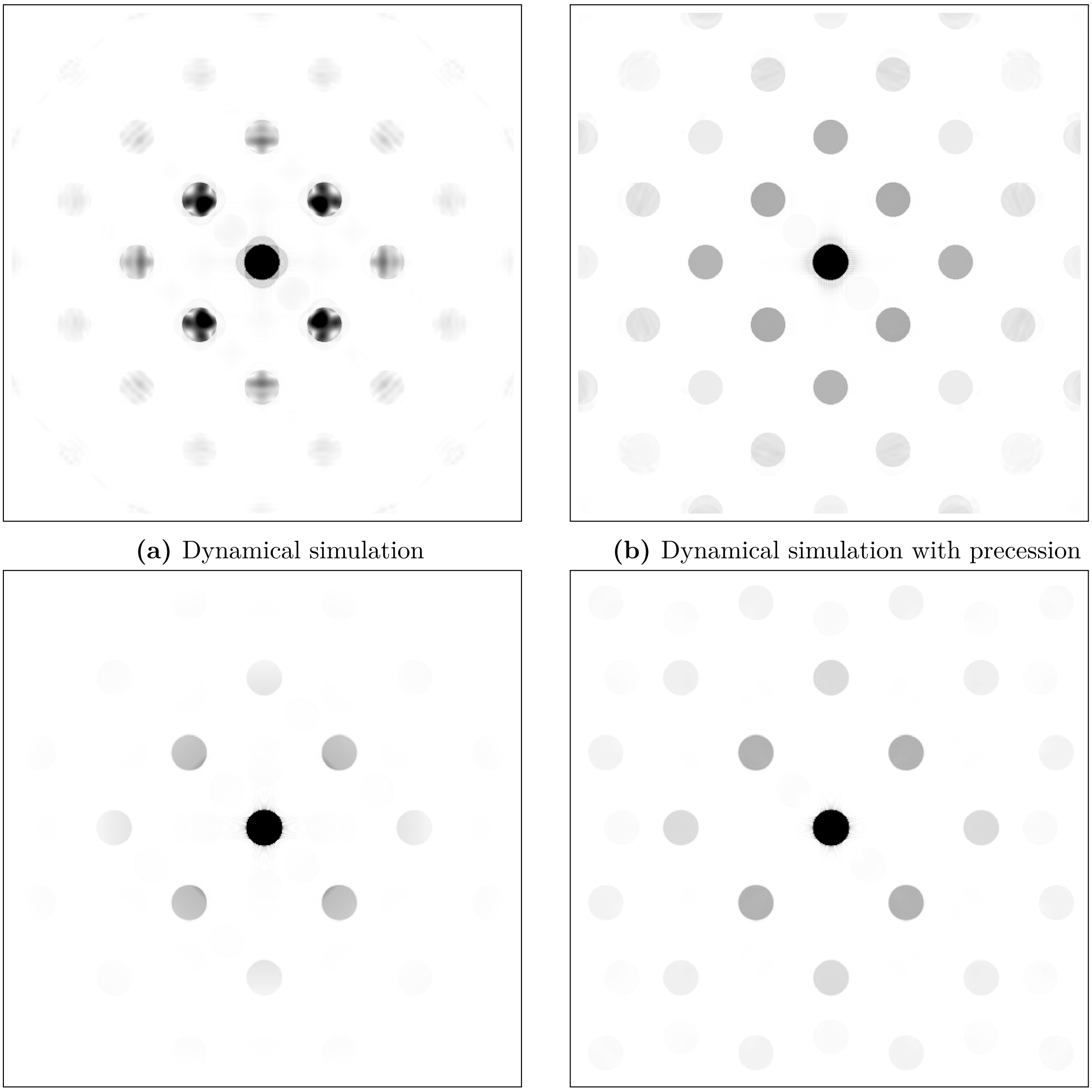


**Figure 1.8** Schematic of double conical beam-rocking geometry used to record precession electron diffraction (PED) patterns. The electron beam is rocked around the optic axis above the sample and counter-rocked below the sample to record electron diffraction patterns containing Bragg disks integrated over the rocking condition.



**(a)** Dynamical simulation

**(b)** Dynamical simulation with precession

**(c)** Kinematical simulation

**(d)** Kinematical simulation with precession

**Figure 1.9** Simulations of diffraction patterns from an unstrained Silicon crystal. (a) shows a dynamical simulation where complex spot inhomogeneities can be seen. (b) shows that with precession, the intensities in the multislice simulation become much more homogeneous. (c)/(d) show kinematical simulations without/with precession. Note that precessed images qualitatively agree very closely with each other.

### 1.4.6 X-ray transform

Section 1.4.3 described three popular methods for simulating diffraction patterns (the forward model), however, these are too computationally complex for use in large scale inverse problems. If a dataset consists of order $10^6$ probe positions, it is more important to have a model which can quickly simulate a full dataset; the full spectral information of a diffraction pattern is not necessary. The common model chosen to fulfil this role is the X-ray transform however, interestingly, different applications use different parts of the microscope to achieve this model. In biological applications it is most common to use bright-field imaging (counting the electrons which pass straight through) whereas in the physical sciences they use dark-field (counting the electrons which diffract). The validity of this, and other large scale linearisations, has been of great interest since around the turn of the millennium and is reviewed by Leary and Midgley (2019). Fundamentally, little trust is put in simulation studies and a new proposed model must be validated with experimental studies.

#### Definition of the X-ray transform

The X-ray transform was first considered by John (1938) and is defined $\mathcal{R}\colon \mathcal{M}(\mathbb{R}^d) \to \mathcal{M}(T\mathbb{S}^{d-1})$ by

$$\mathcal{R}[u](\boldsymbol{\theta}, \boldsymbol{x}) = \int_{\mathbb{R}} u(\boldsymbol{x} + t\boldsymbol{\theta}) dt,$$

where

$$\begin{aligned} \mathbb{S}^{d-1} &= \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \text{ s.t. } |\boldsymbol{\theta}| = 1 \right\}, \\ \text{and } T\mathbb{S}^{d-1} &= \left\{ (\boldsymbol{\theta}, \boldsymbol{x}) \in \mathbb{S}^{d-1} \times \mathbb{R}^d \text{ s.t. } \boldsymbol{\theta} \bullet \boldsymbol{x} = 0 \right\}. \end{aligned}$$

Figure 1.10 provides a schematic of the X-ray transform in 2D. Formally, $T\mathbb{S}^{d-1}$ is just the tangent bundle to the sphere, however, we identify it as the set of all lines in $\mathbb{R}^d$ under the mapping $(\boldsymbol{\theta}, \boldsymbol{x}) \mapsto \{\boldsymbol{x} + t\boldsymbol{\theta} \text{ s.t. } t \in \mathbb{R}\}$. The definition of the range of the X-ray transform is important when asking whether the transform is surjective/invertible. The issue lies in the fact that the current definition parametrises lines non-uniquely, for example $\mathcal{R}u(-\boldsymbol{\theta}, \boldsymbol{x}) = \mathcal{R}u(\boldsymbol{\theta}, \boldsymbol{x})$ for all $(\boldsymbol{\theta}, \boldsymbol{x}) \in T\mathbb{S}^{d-1}$. Fortunately, this is the only non-uniqueness and we will see in the Section Fourier slice theorem that the X-ray transform is indeed invertible on its range.

When $d = 2$, the X-ray transform coincides with the Radon transform (Radon, 1917),

$$\widetilde{\mathcal{R}}u(\boldsymbol{\theta}, \boldsymbol{x}) = \int_{(\boldsymbol{y}-\boldsymbol{x}) \bullet \boldsymbol{\theta} = 0} u(\boldsymbol{y}) d\boldsymbol{y}$$

with $\mathcal{R}u(\boldsymbol{\theta}, \boldsymbol{x}) = \widetilde{\mathcal{R}}u(\boldsymbol{\theta}^\perp, \boldsymbol{x})$ and so the two names are often used interchangeably. In general there is no equivalence, the X-ray transform is the integral over one-dimensional lines and the Radon transform is the integral over one co-dimensional hyperplanes.

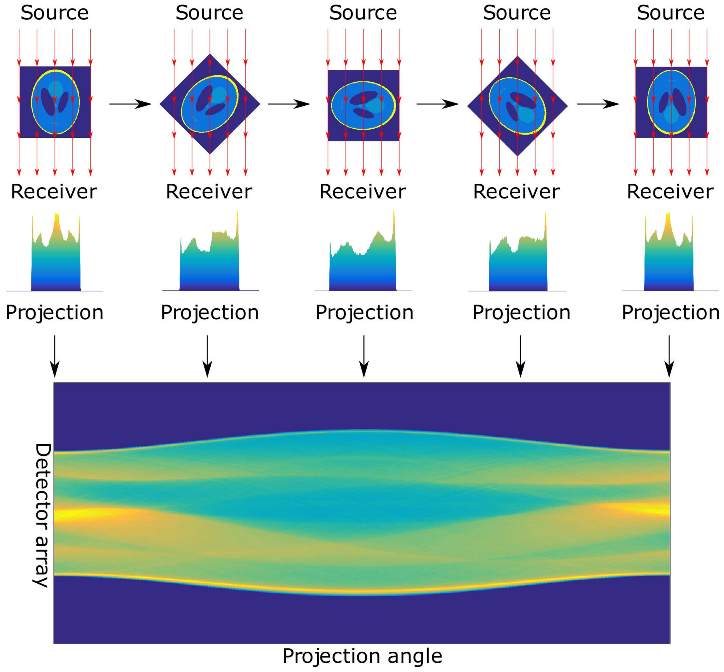


**Figure 1.10** Schematic of the X-ray transform. Electron beams are scanned through the sample and collected at the receiver as the sample is rotated. After 180° rotation the data is a reflected copy of the initial projection.

The definition of the X-ray transform has been extended to a general manifold setting (called the *geodesic X-ray transform*) integrating over any dimensional sub-spaces, see for instance Helgason (1980), however this generality will not be needed in this work.

### Derivation from the Ewald sphere model

To use the X-ray transform in electron microscopy, we want to link it with one of the physical models considered in Section 1.4.3. The Ewald sphere model is a clear choice for this purpose. If we look at the direct beam (electrons which are not diffracted), then (1.5) simplifies to

$$D(0) = |\mathcal{F}[\Psi_p V]\,(0)|^2 = \left|\int_{\mathbb{R}^3} \Psi_p(\boldsymbol{r})V(\boldsymbol{r})\exp(0)d\boldsymbol{r}\right|^2.$$

If we further assume that the electron beam is very narrow, i.e.

$$\Psi_p(\boldsymbol{r}) = \begin{cases} \frac{1}{\pi\varepsilon^2} & r_x^2+r_y^2 \leq \varepsilon^2 \\ 0 & \text{else} \end{cases}$$

for small $\varepsilon$, then this becomes

$$\sqrt{D(0)} = \int_{\mathbb{R}} \fint_{r_x^2+r_y^2\leq\varepsilon^2} V(\boldsymbol{r})dr_x dr_y dr_z \sim \mathcal{R}[V]\left(\left(\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right), \left(\begin{smallmatrix}0\\0\\0\end{smallmatrix}\right)\right).$$

This argument validates the use of the X-ray transform for both dark- and light-field imaging.

### Heuristic derivation

The Ewald sphere model is one way to justify the use of the X-ray transform in EM but we can also provide a much more heuristic argument to suggest that the X-ray transform is the only natural linearisation.

Let $I\in\mathbb{R}$ be the observed intensity from a single electron beam. We are attempting to build a model to predict the value of $I$ from physical parameters:

1. If the model is linear with respect to the scattering potential $V$, then there exists a kernel $\varphi$ such that
$$I=\int_{\mathbb{R}^3}\varphi(\boldsymbol{r})V(\boldsymbol{r})d\boldsymbol{r}.$$
2. If the electron is a particle which travels on curve $\vec{\gamma}$ and $I$ depends only on the values of $V$ on the electron path, then
$$I=\int_{\mathbb{R}}\varphi(\vec{\gamma}(t))V(\vec{\gamma}(t))dt.$$
3. If un-diffracted electrons travel in straight lines, then $\vec{\gamma}(t)=(0,0,t)$
$$I=\int_{\mathbb{R}}\varphi(0,0,t)V(0,0,t)dt.$$
4. If $I$ is insensitive to movement in the $z$-direction, then $\varphi$ is scalar
$$I=\int_{\mathbb{R}}\varphi V(0,0,t)dt=\varphi\mathcal{R}[V]\left(\left(\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right),\left(\begin{smallmatrix}0\\0\\0\end{smallmatrix}\right)\right).$$

It is interesting to consider what changes when certain assumptions are weakened. Assumption (1) is the most fragile. As was seen in Section 1.4.4, the crystalline properties of a sample change the diffraction behaviour dramatically and it is unlikely that this information can be compressed into a grey-scale potential $V$. The other factor is the Beer-Lambert law (Levine, 2005) which states that the beam intensity should decay exponentially, this corresponds to the attenuated X-ray transform. The attenuated X-ray transform is a harder problem but still well studied (Novikov, 2002; Fokas et al., 2005).

Moving on to (2) and (3), if the electron does not travel on a 1D line, then this can be incorporated as a blurring operator. If the lines are not straight, then this is simply the case of the geodesic X-ray transform. Finally, for point (4), the electron beam is focussed on a particular $z$-plane. Anything above/below that plane is likely to interact with the beam differently.

Despite these caveats, the X-ray transform is currently the best known forward model used to find 3D reconstructions of samples in electron microscopy. The generality of this argument may also go some way to explain the prevalence of the X-ray transform in many diverse applications and modalities. As already discussed, it is the main model in electron tomography in both physical and biological fields (Leary et al., 2013; Kübel et al., 2005; Zhao et al., 2013). As the name suggests, it is also the model used in many X-ray imaging applications including medical CT (Kalender, 2006) and geosciences (Cnudde and Boone, 2013). Another common imaging modality is Positron Emission Tomography (PET) where the counting of positrons is also modelled with the X-ray transform. Each modality comes with its own derivation, approximations, and modelling errors however, in principle, any methodology developed in one modality is immediately applicable in the others.

### Fourier slice theorem

Arguably the biggest single advancement in the analysis and application of the X-ray transform is realising its connection to the Fourier transform. First discovered by Bracewell (1956), for the X-ray transform it can be stated as follows.

**Theorem 1.4.4** (Solmon (1976))**.** *If $u \in L^1(\mathbb{R}^d)$, then*

$$\mathcal{F}_{d-1}[\mathcal{R}u(\boldsymbol{\theta},\cdot)](\boldsymbol{k}) = \mathcal{F}_d[u](\boldsymbol{k})$$

*for all $|\boldsymbol{\theta}| = 1$ and $\boldsymbol{k} \perp \boldsymbol{\theta}$ where $\mathcal{F}_{d-1}$ is the $d-1$-dimensional Fourier transform over $\boldsymbol{\theta}^\perp$ and $\mathcal{F}_d$ is the $d$-dimensional Fourier transform over $\mathbb{R}^d$.*

*Proof.* We proceed by definition,

$$\begin{aligned}\mathcal{F}_{d-1}[\mathcal{R}u(\boldsymbol{\theta},\cdot)](\boldsymbol{k}) &= \int_{\boldsymbol{\theta}^\perp}\int_{\mathbb{R}} u(\boldsymbol{x}+t\boldsymbol{\theta})\exp(-\imath\boldsymbol{x}\cdot\boldsymbol{k})dtd\boldsymbol{x} \\ &= \int_{\boldsymbol{\theta}^\perp}\int_{\mathbb{R}} u(\boldsymbol{x}+t\boldsymbol{\theta})\exp(-\imath(\boldsymbol{x}+t\boldsymbol{\theta})\cdot\boldsymbol{k})dtd\boldsymbol{x} && \boldsymbol{\theta}\cdot\boldsymbol{k}=0 \\ &= \int_{\mathbb{R}^d} u(\boldsymbol{x})\exp(-\imath\boldsymbol{x}\cdot\boldsymbol{k})d\boldsymbol{x} \\ &= \mathcal{F}[u](\boldsymbol{k}).\end{aligned}$$

As $u \in L^1$, Fubini's theorem guarantees the manipulation of integral domains. □

This is a hugely powerful result because it equates the X-ray transform with pointwise sampling in the Fourier domain, something which is well understood.

### Analytical inversion

A consequence of the Fourier slice theorem is the *filtered back-projection* (FBP) inversion formula.

**Theorem 1.4.5** (Solmon (1976))**.** *If $u\in L^1(\mathbb{R}^d)$, then*

$$u(\boldsymbol{x}) = \mathcal{R}^\top(\varphi\star\mathcal{R}u)(\boldsymbol{x})$$

*for all $\boldsymbol{x}\in\mathbb{R}^3$ where*

$$\varphi\colon\mathbb{R}^d\to\mathbb{R},\quad \varphi(\boldsymbol{x}) = |\boldsymbol{x}|^{1-d}.$$

*Equivalently, $\mathcal{F}_{d-1}[\varphi](\boldsymbol{k}) = \alpha_d|\boldsymbol{k}|^{-1}$ where $\alpha_d = \pi^{\frac{d}{2}}2^{d-1}\frac{\Gamma(d/2)}{\Gamma(1/2)}$.*

*Proof.* First we expand,

$$\begin{aligned}\mathcal{R}^\top\mathcal{R}u(\boldsymbol{x}) &= \int_{\mathbb{S}^{d-1}}\mathcal{R}u(\boldsymbol{\theta},\boldsymbol{x})d\boldsymbol{\theta}\\ &= \int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}}u(\boldsymbol{x}+t\boldsymbol{\theta})dtd\boldsymbol{\theta}\\ &= \int_{\mathbb{R}^d}u(\boldsymbol{x}-\boldsymbol{y})\frac{d\boldsymbol{y}}{\beta_{d-1}|\boldsymbol{y}|^{d-1}}\end{aligned}$$

where $\beta_{d-1}$ is the change of basis scaling constant:

$$\begin{aligned}\text{volume}(\mathbb{S}^{d-1}) = \int_{|\boldsymbol{x}|\leq 1}1d\boldsymbol{x} = \int_{\mathbb{S}^{d-1}}\int_0^1\beta_{d-1}|t|^{d-1}dtd\boldsymbol{\theta} = \frac{\text{surface area}(\mathbb{S}^{d-1})}{d}\\ \rightsquigarrow \beta_{d-1} = \frac{1}{d}\frac{\text{surface area}(\mathbb{S}^{d-1})}{\text{volume}(\mathbb{S}^{d-1})} = 1.\end{aligned}$$

Combining this with the properties of Theorem 1.4.4 shows that $\mathcal{R}$ commutes with convolution in the required way. The formula

$$\mathcal{F}_d[|\boldsymbol{y}|^{1-d}] = \alpha_d|\boldsymbol{k}|^{-1}$$

can be found in a standard lookup table. □

The form of this inverse confirms that the fully sampled X-ray transform is invertible but not well posed. If we are finding the least-squares solution to $\mathcal{R}u=\eta$ where $\eta = \mathcal{R}u^\dagger+\varepsilon$ for some noise $\varepsilon$ with Fourier frequency $k$, then we get

$$u = (\mathcal{R}^\top\mathcal{R})^{-1}\mathcal{R}^\top(\mathcal{R}u^\dagger+\varepsilon) = u^\dagger + O(\sqrt{k}\,\|\varepsilon\|_2).$$

This is not a bad sensitivity to noise but does demonstrate that high frequency noise is amplified by naive reconstructions. The standard approach is to modify the convolution kernel $\varphi$ such that it dampens the high-frequency components based on an estimation of the noise level (Lyra and Ploussi, 2011). For example:

$$\text{Ramp filter: } \mathcal{F}[\varphi] \propto \begin{cases} |\boldsymbol{k}| & |\boldsymbol{k}| \leq k_c \\ 0 & \text{else} \end{cases},$$

$$\text{Hanning filter: } \mathcal{F}[\varphi] \propto \begin{cases} \left[1 + \cos\left(\pi \frac{|\boldsymbol{k}|}{k_c}\right)\right] |\boldsymbol{k}| & |\boldsymbol{k}| \leq k_c \\ 0 & \text{else} \end{cases}.$$

Both of these are commonly used in applications. We can also extend the inversion formula to a sub-sampled pseudo-inverse.

**Theorem 1.4.6** (Solmon (1976))**.** *If $\mathcal{A} = \mathcal{R}|_{\boldsymbol{\theta}\in\Theta}$ for some set $\Theta \subset \mathbb{S}^{d-1}$, then*

$$\mathcal{A}\mathcal{A}^\top(\varphi \star \mathcal{A}u)(\boldsymbol{x}) = \mathcal{A}u$$

*for all $u \in L^1(\mathbb{R}^3)$ and $\varphi$ is as in Theorem 1.4.5.*

*Proof.* Following the previous proof,

$$\begin{aligned} \mathcal{A}^\top\mathcal{A}u(\boldsymbol{x}) &= \sum_{\boldsymbol{\theta}\in\Theta} \mathcal{A}u(\boldsymbol{\theta}, \boldsymbol{x}) \\ &= \int_\Theta \int_{\mathbb{R}} u(\boldsymbol{x} + t\boldsymbol{\theta}) dt d\boldsymbol{\theta} \\ &= \int_{\mathbb{R}^d} u(\boldsymbol{x} - \boldsymbol{y}) \int_\Theta \delta_0(\boldsymbol{y} - (\boldsymbol{y} \cdot \boldsymbol{\theta})\boldsymbol{\theta}) d\boldsymbol{\theta} \frac{d\boldsymbol{y}}{|\boldsymbol{y}|^{d-1}}. \end{aligned}$$

The only additional complexity is the Dirac function. If we apply the inversion kernel slice-wise on the observed data, then the Dirac is ignored and the scaling removed as before. If we apply $\mathcal{A}$ again, then the Dirac functions are naturally absorbed. □

This argument allows us to extend the FBP to a pseudo-inverse for other datasets with missing angular information and finally understand the reconstructions in Figure 1.4, each performed with the FBP. For convenience, we replicate the plots of Figure 1.4 in Figure 1.11.

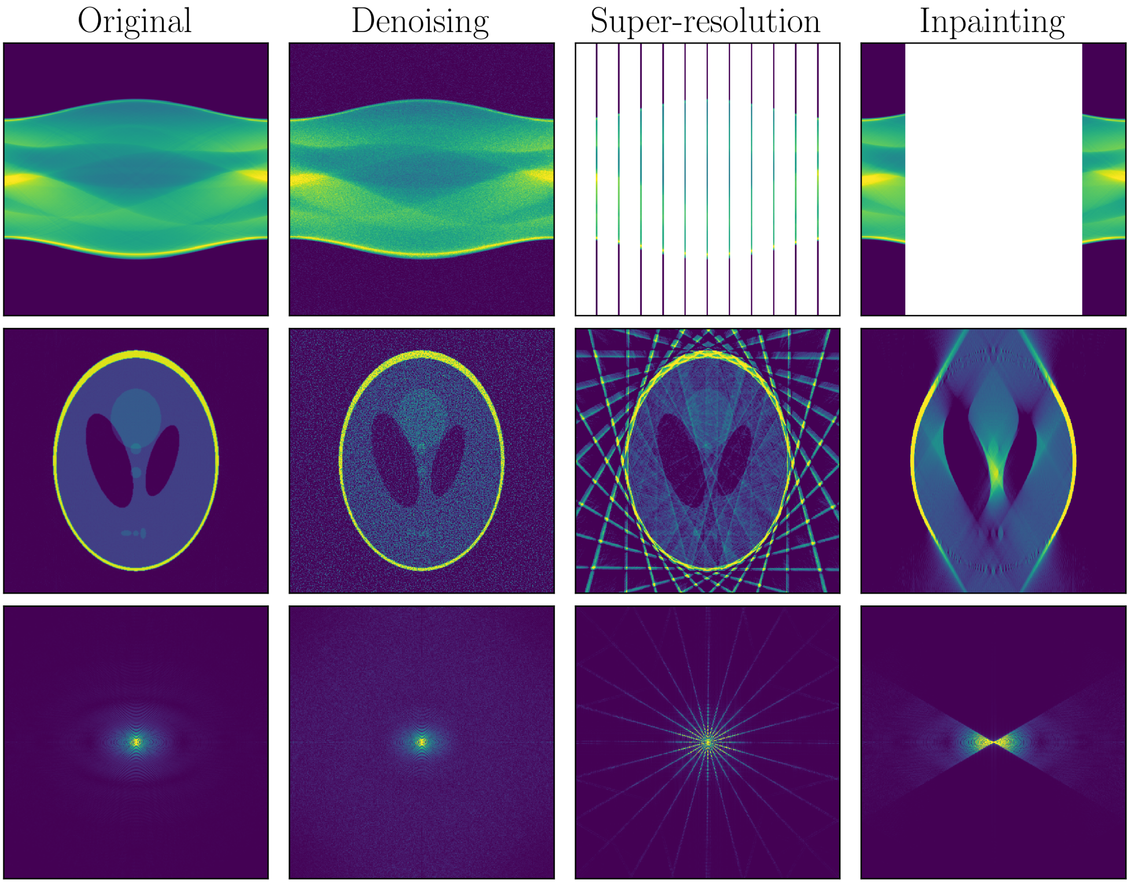


**Figure 1.11** Extension of Figure 1.4. Filtered back reconstructions from various corrupted datasets. Top row shows raw data, middle shows FBP reconstruction, third row shows discrete Fourier transform of reconstruction.

The original reconstruction from high-resolution noise free data (first column) is very accurate. If we keep the same resolution but add noise, then the resulting reconstruction amplifies that noise (second column). When we reduce the angular resolution of the data (super-resolution task), FBP introduces the characteristic streak artifacts. This can be seen in the analytical formula as $u^* = u^\dagger \star [\sum_{\boldsymbol{\theta}\in\Theta} \delta_0(\boldsymbol{y} - (\boldsymbol{y}\cdot\boldsymbol{\theta})\boldsymbol{\theta})]$. Finally, when a region of X-ray coefficients are not recorded in the data, the reconstruction formula becomes $u^* = u^\dagger \star \left[\mathbb{1}_\Theta(\frac{\boldsymbol{y}}{|\boldsymbol{y}|})\right]$. The characteristics of each reconstruction method can also be seen in the Fourier transforms of the reconstructions. The presence of noise adds a background intensity in the Fourier domain and Fourier coefficients are assumed to be 0 if they are not present in the original data.

### 1.4.7 Limited angle tomography

Limited angle tomography is very common in EM and forms one of the foci of this thesis. In theoretical terms, this is the inpainting scenario depicted in Figure 1.12 where a large range of

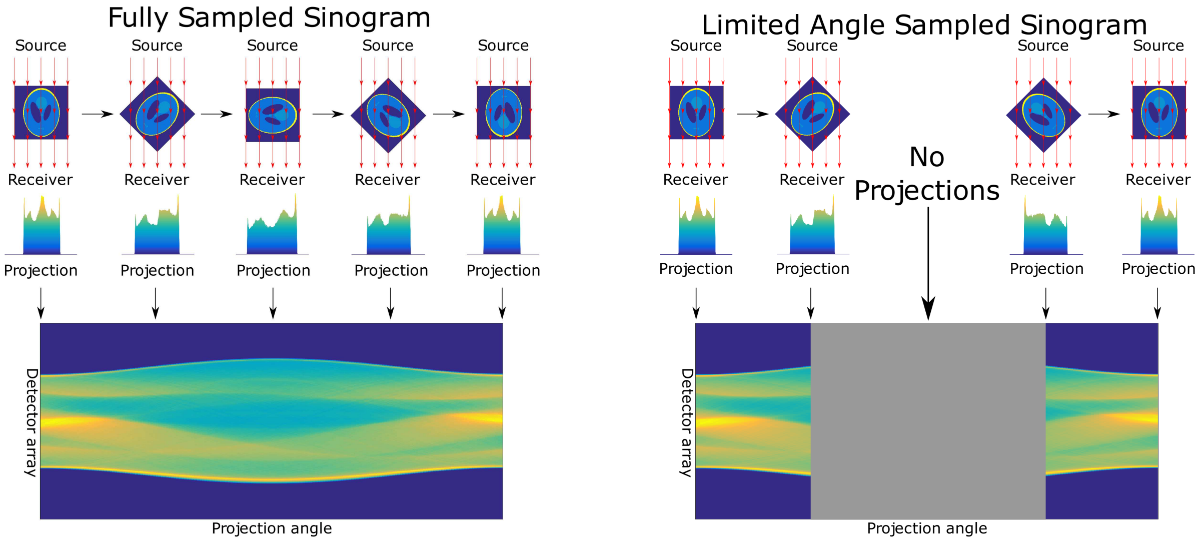


**Figure 1.12** Schematic of the acquisition of 2D X-ray transform data , the sinogram, in both full range and limited angle acquisition. Note that measurement at 180° is exactly a reflection of that at 0°. This symmetry allows us to consider a 180° range of the sinogram as a full sample. In the limited angle setting we can only rotate the sample a small amount clockwise and anti-clockwise which results in missing data in the middle of the sinogram.

angular information is missing from the observations. The analytical form is

$$\mathcal{A}\colon \mathcal{M}(\mathbb{R}^d) \to T\mathbb{S}^{d-1}, \qquad \mathcal{A} = \mathcal{S}\mathcal{R}$$

where $\mathcal{S}\colon \mathcal{M}(T\mathbb{S}^{d-1}) \to \mathcal{M}(T\Theta)$ is the canonical projection and $\Theta \subset \mathbb{S}^{d-1}$ is a simply connected set. Through the Fourier slice theorem, we can view this as an inpainting problem in the Fourier domain. Fourier coefficients are observed everywhere other than the so-called *missing wedge* (or *missing cone* in 3D). An example of this has already been shown in the right-hand column of Figure 1.11 where we clearly see this missing wedge in the Fourier domain of the reconstruction. In practical terms, the limited angle scenario arises from having a limited range of possible tilts, as motivated in Figure 1.12. From an engineering perspective, specimen holders can typically allow up to ±70° of rotation (77 % of full data) before the sample is obscured by parts of the specimen holder. This forms a lower bound on the missing wedge, however, other considerations increase the wedge in order to ensure the validity of the X-ray transform, as is well described by Leary and Midgley (2019). Two examples are demonstrated in Figure 1.13, one of which is that if the tilt becomes too large, then the thin-specimen assumption may be violated and this may push the model into a non-linear regime. In practice, especially in biological EM, it is still common to be limited to between 30° to 40° (33 % to 44 % of full data), for instance in Vilas et al. (2020).

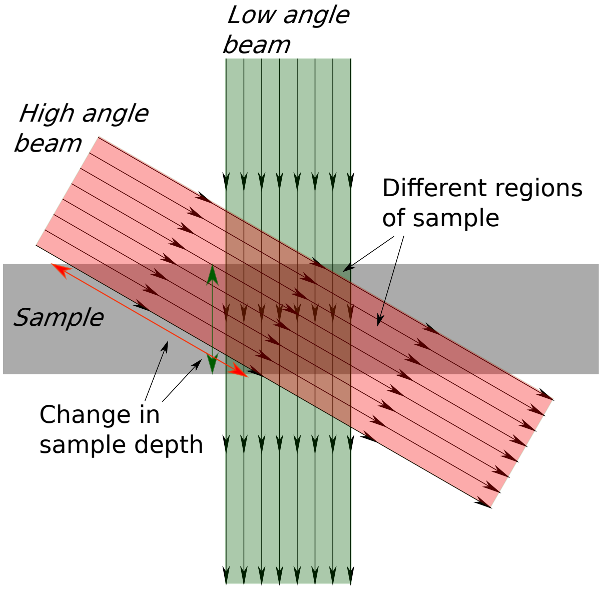


**Figure 1.13** Schematic of problems with large tilts, visualised with the beam rotating and the specimen stationary. The primary issue is that samples are typically large and thin. When this is tilted the thickness of the sample, relative to the electron beam, increases rapidly. The X-ray transform also assumes that everything is visible in every tilt. If the sample is larger than the detector, then different parts of the sample will be visible at different tilts. Again, this worsens at large tilts.

As has been motivated in Section 1.2.1, inpainting is a very challenging task in inverse problems and reconstruction errors are often observed. Two common methods for reconstruction of corrupted tomography data in EM are:

- FBP, as introduced in Theorem 1.4.5.
- Simultaneous Iterative Reconstruction Technique (SIRT), which is a preconditioned gradient descent on the function $\|\mathcal{A}u - \eta\|_2^2$ with early stopping to reduce the effect of noise (Gilbert, 1972; Kübel et al., 2005; Agulleiro et al., 2010; Spitzbarth and Drescher, 2015).
- TV reconstruction, variational reconstruction with the energy $\mathrm{E}(u) = \frac{1}{2}\|\mathcal{A}u - \eta\|_2^2 + \mu\,\mathrm{TV}(u)$ (Goris et al., 2012; Leary et al., 2013; Collins et al., 2017).

In broad strokes: FBP works well when the data is very good, SIRT works well even when there is some noise, TV reconstructions perform well even when some data is missing. Figure 1.14 gives some examples of reconstructions from moderately bad data which demonstrates this point in examples of noisy super-resolution and noiseless inpainting. An intriguing observation is that the TV reconstruction with noise-free data is perfect (with $\mu \to 0$), even with only 33% of the data in an inpainting scenario. Clean data is not guaranteed to give a perfect reconstruction,

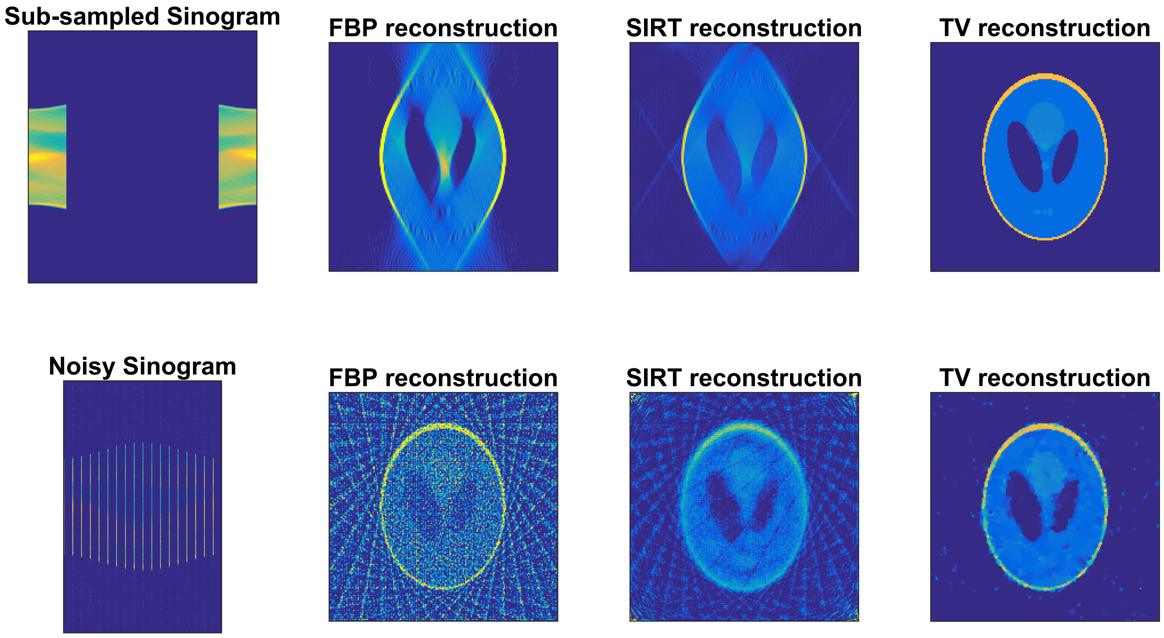


**Figure 1.14** Demonstration of TV reconstruction in comparison to FBP and SIRT. The top row shows reconstructions from noiseless limited angle data and the bottom shows reconstructions from noisy limited view data (far left images). Comparing the columns, we immediately see that FBP and SIRT are much more prone to angular artifacts than TV. In both cases we notice that the TV reconstructions better show the broad structure of the phantom.

for instance in Figure 1.15, however, the results are very reasonable. This behaviour changes as soon as there is noise, as demonstrated in Figure 1.16. Interestingly, it is not the noise itself which causes the elongation but the choice of $\mu$. There are three scenarios which arise:

$$\begin{aligned}
&\text{no noise, } \mu = 0 && \rightsquigarrow u^* = \operatorname{argmin}\left\{\mathrm{TV}(u) \text{ s. t. } \mathcal{A}u = \mathcal{A}u^\dagger\right\} && u^* \text{ is almost perfect}\\
&\text{no noise, } \mu > 0 && \rightsquigarrow u^* = \operatorname{argmin} \tfrac{1}{2}\left\|\mathcal{A}u - \mathcal{A}u^\dagger\right\|_2^2 + \mu\,\mathrm{TV}(u) && \text{with elongation artifacts}\\
&\text{with noise, } \mu > 0 && \rightsquigarrow u^* = \operatorname{argmin} \tfrac{1}{2}\left\|\mathcal{A}u - \eta\right\|_2^2 + \mu\,\mathrm{TV}(u) && \text{with noise and elongation}
\end{aligned}$$

The first case is shown in Figure 1.14 and the last in Figure 1.16. From an inverse problems perspective, this demonstrates the failure of TV as a good regulariser for limited angle reconstructions. Considering the criteria for a well-posed inverse problem, the reconstruction may be continuous with respect to $\mu$ but is much less stable in a limited angle scenario. Despite this relatively clear failure, TV remains one of the best reconstruction methods available for limited angle problems as it is hard to find a categorically better method. An attempt to address this problem will be proposed in Chapter 2.

The full mathematical analysis of the limited angle problem is described by tools in microlocal analysis (Quinto, 1993; Krishnan and Quinto, 2015) which allow formal expression of what information of $u^\dagger$ is lost when Fourier coefficients are missing (Frikel and Quinto, 2013; Katsevich, 1997). The core principle is that every sample is a sum of singularities and can be partitioned into *visible* and *invisible singularities* depending on whether they are 'visible' in the sampled data or not. For the X-ray transform, the partition between visible/invisible singularities corresponds exactly with the Fourier domain images in Figure 1.11. In the context of Figures 1.14 to 1.16 (i.e. when $u^\dagger$ is piecewise constant and the missing angles are centred at $\theta = 90°$), then the location of the near-horizontal edges is missing from the data. Each reconstruction method attempts to 'guess' the location of those edges in a particular way. The FBP tries to assert that there are *no* horizontal edges whereas TV asserts that $u^\dagger$ is 'smooth' in the vertical direction. The heuristic of TV was that smoothness should equate to sparse gradient, however, in this instance we see a failing of the convex approximation of sparsity.

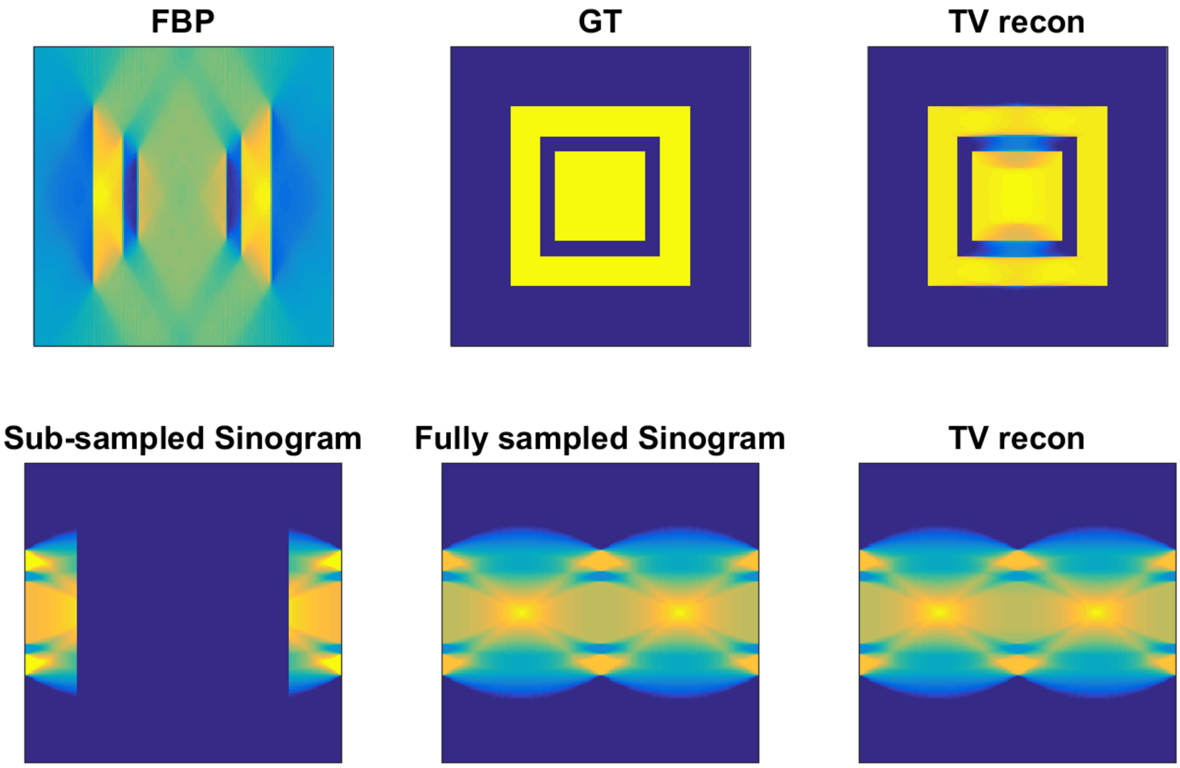


**Figure 1.15** Example where TV reconstruction is not perfect with clean limited angle data and $\mu = 0$.

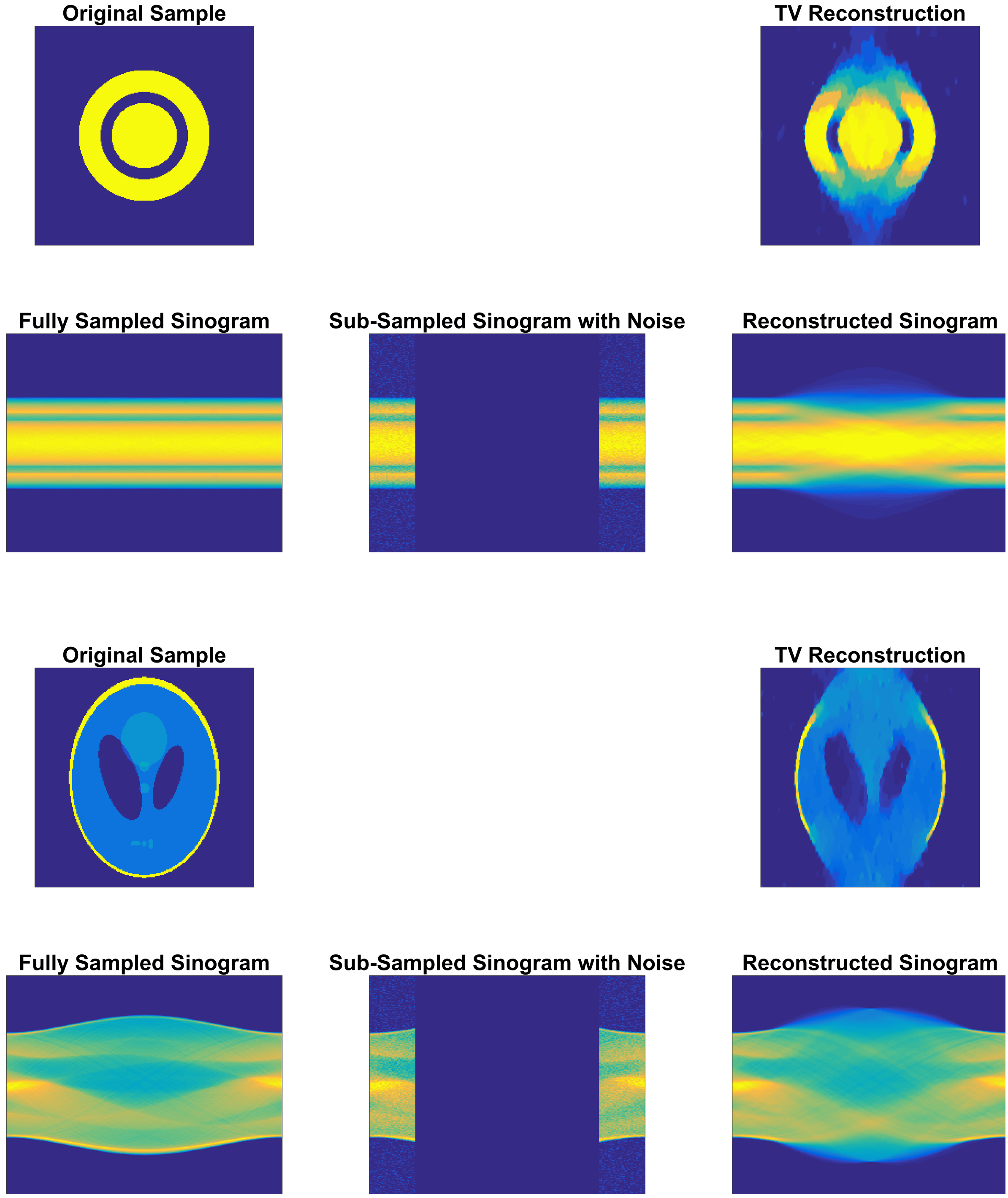


**Figure 1.16** Examples when TV reconstructions cannot recover the global structures of samples from noisy data and $\mu > 0$. When there is a large missing wedge ($^2\!/_3$ of data unseen) and noise on the projections, then reconstructions exhibit characteristic blurring in the vertical direction. This can also be seen in the extrapolated region of the sinograms as a loss of structure.

### 1.4.8 Tensor tomography

The standard X-ray transform defined in Section 1.4.6 was for reconstructing greyscale objects from greyscale images. For richer physical structures this needs to be generalised. There are many generalisations in the Euclidean (Sharafutdinov, 1994) and manifold settings (Paternain et al., 2014), however, we will focus on the Euclidean domain $\mathbb{R}^3$ for simplicity of notation and to cover the physically relevant case.

**Definition 1.4.7.** *If $\Omega \subset \mathbb{R}^3$, then we define the* Longitudinal Ray Transform *(LRT)* $\mathrm{LRT}\colon L^1(\Omega, \mathbb{R}^3) \to T\mathbb{S}^2$ *for $u \in L^1(\Omega, \mathbb{R}^3)$ by*

$$\mathrm{LRT}[\vec{u}](\boldsymbol{\theta}, \boldsymbol{x}) = \int_{\mathbb{R}} \vec{u}(\boldsymbol{x} + t\boldsymbol{\theta}) \cdot \boldsymbol{\theta} dt$$

*and the* Transverse Ray Transform *(TRT)* $\mathrm{TRT}\colon L^1(\Omega, \mathbb{R}^{3\times 3}) \to T\mathbb{S}^2 \times \mathbb{R}^{3\times 3}$ *by*

$$\mathrm{TRT}[\vec{U}](\boldsymbol{\theta}, \boldsymbol{x}) = \int_{\mathbb{R}} \Pi_{\boldsymbol{\theta}} \vec{U}(\boldsymbol{x} + t\boldsymbol{\theta}) \Pi_{\boldsymbol{\theta}} dt, \qquad \Pi_{\boldsymbol{\theta}} = \mathrm{id} - \boldsymbol{\theta}\boldsymbol{\theta}^\top.$$

The remainder of this subsection is dedicated to computing the null-spaces of the LRT and TRT in three dimensions. Most of the analysis is already well understood in an arbitrary number of dimensions, for instance Sharafutdinov (1994) showed that the LRT is never invertible and it was later shown that the TRT is invertible in dimensions strictly greater than three whenever $\Omega$ is a simple real analytic Riemannian manifold (Novikov and Sharafutdinov, 2007; Abhishek, 2020).

In three dimensions, the TRT is only invertible on the subspace of tensor fields of symmetric tensors (Holman, 2013; Lionheart and Withers, 2015). Holman (2013) computes the null-space of the TRT (complete with non-symmetric tensors) in the manifold setting, but it is hard to find a similarly self-contained argument in the much simpler Euclidean case.

Computations are typically performed in the Schwartz space, $\mathcal{S}(\mathbb{R}^3, \mathbb{R}^3)$ or $\mathcal{S}(\mathbb{R}^3, \mathbb{R}^{3\times 3})$, where functions have infinite smoothness in the real and Fourier domains. Recent work by Boman and Sharafutdinov (2018) has extended the stability results of the LRT to Sobolev spaces on compact domains and the same techniques appear to be valid for the TRT but this has not yet been made rigorous.

Sticking to the classical Schwartz space setting, the null-spaces of the LRT and symmetric TRT are computed by Sharafutdinov (1994); Desai and Lionheart (2016) respectively. Their results are summarised in the following theorem.

**Theorem 1.4.8** ((Sharafutdinov, 1994, Chapter 2),(Desai and Lionheart, 2016, Theorem 1))**.** *For all $\vec{u}, \vec{v} \in \mathcal{S}(\mathbb{R}^3, \mathbb{R}^3)$, $\vec{U}, \vec{V} \in \mathcal{S}(\mathbb{R}^3, \mathrm{Sym}(\mathbb{R}^{3\times 3}))$, orthogonal bases $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$:*

$$\begin{aligned}
\mathrm{LRT}[\vec{u}](\boldsymbol{\theta},\boldsymbol{x}) = \mathrm{LRT}[\vec{v}](\boldsymbol{\theta},\boldsymbol{x}) \quad & \forall(\boldsymbol{\theta},\boldsymbol{x}) \in T\$^2 && \iff && \vec{u} = \vec{v} + \nabla\varphi \textit{ for some } \varphi \in \mathcal{S}(\mathbb{R}^3),\\
\mathrm{TRT}[\vec{U}](\boldsymbol{\theta},\boldsymbol{x}) = \mathrm{TRT}[\vec{V}](\boldsymbol{\theta},\boldsymbol{x}) \quad & \forall\boldsymbol{\theta} \in \cup\boldsymbol{e}_i^\perp,\ \boldsymbol{x} \perp \boldsymbol{\theta} && \iff && \vec{U} = \vec{V}.
\end{aligned}$$

In other words, the TRT is injective on the domain of tensor fields of symmetric tensors if data is collected from three tilt series but the LRT always has a kernel of conservative gradient fields.

We now finally move on to the null-space computation for the TRT for tensor fields of non-symmetric tensors. The case when $\Omega$ is a general simple compact manifold is shown in Holman (2013) (the combination of Theorem 4 and Equation 13) although is much more complicated than needed in the Euclidean case. The following argument closely follows that of (Novikov and Sharafutdinov, 2007, Section 4), however we also retain the symmetric component of the tensor field.

**Theorem 1.4.9.** *For all $\vec{U},\vec{V} \in \mathcal{S}(\mathbb{R}^3,\mathbb{R}^{3\times3})$, orthogonal bases $\{\boldsymbol{e}_1,\boldsymbol{e}_2,\boldsymbol{e}_3\}$, then*

$$\mathrm{TRT}[\vec{U}](\boldsymbol{\theta},\boldsymbol{x}) = \mathrm{TRT}[\vec{V}](\boldsymbol{\theta},\boldsymbol{x}) \quad \forall\boldsymbol{\theta} \in \cup\boldsymbol{e}_i^\perp,\ \boldsymbol{x} \perp \boldsymbol{\theta} \iff \vec{U} = \vec{V} + [\nabla\varphi]_\times \textit{ for some } \varphi \in \mathcal{S}(\mathbb{R}^3)$$

*where* $\mathrm{Sym}(\vec{U})(\boldsymbol{x}) = \frac{1}{2}\left(\vec{U}(\boldsymbol{x}) + \vec{U}(\boldsymbol{x})^\top\right)$.

The proof of Theorem 1.4.9 relies on the following pointwise decomposition lemma. In what follows define $[\boldsymbol{r}]_\times$ to be the matrix such that $[\boldsymbol{r}]_\times \boldsymbol{r}' = \boldsymbol{r} \times \boldsymbol{r}'$ for all $\boldsymbol{r}' \in \mathbb{R}^3$. Similarly, define $\left[\vec{\mathrm{f}}\right]_\times(\boldsymbol{r}) = \left[\vec{\mathrm{f}}(\boldsymbol{r})\right]_\times$ to be the pointwise operation.

**Lemma 1.4.10.** *The TRT of a general tensor field can be decomposed as:*

1. *For all $\vec{U} \in \mathcal{S}(\mathbb{R}^3;\mathbb{R}^{3\times3})$ there exists $\vec{u} \in \mathcal{S}(\mathbb{R}^3;\mathbb{R}^3)$ such that*

$$\vec{U} = \mathrm{Sym}(\vec{U}) + [\vec{u}]_\times$$

2. $\mathrm{TRT}[\mathrm{Sym}(\vec{U})](\boldsymbol{\theta},\boldsymbol{x}) = \mathrm{Sym}(\mathrm{TRT}[\vec{U}](\boldsymbol{\theta},\boldsymbol{x}))$
3. $\mathrm{TRT}\left[[\vec{u}]_\times\right](\boldsymbol{\theta},\boldsymbol{x}) = \mathrm{LRT}\left[\vec{U}(\boldsymbol{\theta},\boldsymbol{x})\right][\boldsymbol{\theta}]_\times$

*Proof of Lemma 1.4.10.* Part *(i)* is a simple algebraic equivalence. For any $A \in \mathbb{R}^{3\times3}$

$$A - \mathrm{Sym}(A) = \frac{1}{2}\begin{pmatrix} 0 & A_{1,2}-A_{2,1} & A_{1,3}-A_{3,1} \\ -(A_{1,2}-A_{2,1}) & 0 & A_{2,3}-A_{3,2} \\ -(A_{1,3}-A_{3,1}) & -(A_{2,3}-A_{3,2}) & 0 \end{pmatrix} = \left[\frac{1}{2}\begin{pmatrix} A_{3,2}-A_{2,3} \\ A_{1,3}-A_{3,1} \\ A_{2,1}-A_{1,2} \end{pmatrix}\right]_\times$$

The expression $\vec{U} = \mathrm{Sym}(\vec{U}) + [\vec{u}]_\times$ is simply the pointwise extension of this equality. Confirming $\vec{u} \in \mathcal{S}$ is also clear as $[\vec{u}]_\times = \frac{1}{2}\vec{U} - \frac{1}{2}\vec{U}^\top \in \mathcal{S}$ and so we must have $\vec{u}_i \in \mathcal{S}$ for each $i = 1,2,3$.

By the linearity of the TRT,

$$\mathrm{TRT}\left[\mathrm{Sym}(\vec{U})\right] = \mathrm{TRT}\left[\tfrac{1}{2}(\vec{U}+\vec{U}^\top)\right] = \tfrac{1}{2}\left(\mathrm{TRT}\left[\vec{U}\right] + \mathrm{TRT}\left[\vec{U}^\top\right]\right).$$

Hence, to prove part *(ii)*, it suffices to show:

$$\mathrm{TRT}\left[\vec{U}^\top\right](\boldsymbol{\theta},\boldsymbol{x}) = \int_{\mathbb{R}} \Pi_{\boldsymbol{\theta}}\vec{U}(\boldsymbol{x}+t\boldsymbol{\theta})^\top \Pi_{\boldsymbol{\theta}} dt = \int_{\mathbb{R}} (\Pi_{\boldsymbol{\theta}}\vec{U}(\boldsymbol{x}+t\boldsymbol{\theta})\Pi_{\boldsymbol{\theta}})^\top dt = \mathrm{TRT}\left[\vec{U}\right](\boldsymbol{\theta},\boldsymbol{x})^\top.$$

The proof of part *(iii)* is also by direct evaluation, fix $\boldsymbol{a}\in\mathbb{R}^3$. We claim $\Pi_{\boldsymbol{\xi}}\left[\boldsymbol{a}\right]_\times \Pi_{\boldsymbol{\theta}} = \langle \boldsymbol{a},\ \boldsymbol{\theta}\rangle \left[\boldsymbol{\theta}\right]_\times$. First note the trivial case, when $\boldsymbol{\theta} = \boldsymbol{e}_3 = \left(\begin{smallmatrix} 0 & 0 & 1\end{smallmatrix}\right)^\top$

$$\Pi_{\boldsymbol{e}_3}\left[\boldsymbol{a}\right]_\times \Pi_{\boldsymbol{e}_3} = \left(\begin{smallmatrix} 1&0&0\\0&1&0\\0&0&0\end{smallmatrix}\right)\begin{pmatrix} 0 & -a_3 & a_2\\ a_3 & 0 & -a_1\\ -a_2 & a_1 & 0\end{pmatrix}\left(\begin{smallmatrix} 1&0&0\\0&1&0\\0&0&0\end{smallmatrix}\right) = \begin{pmatrix} 0 & -a_3 & 0\\ a_3 & 0 & 0\\ 0 & 0 & 0\end{pmatrix} = \left[a_3\boldsymbol{e}_3\right]_\times.$$

To generalise this, recall the property of cross products $(R\boldsymbol{a})\times(R\boldsymbol{b}) = R(\boldsymbol{a}\times\boldsymbol{b})$ for all rotation matrices $R$ and vectors $\boldsymbol{b}$. Because of this, we know

$$R\left[\boldsymbol{a}\right]_\times R^\top\boldsymbol{b} = R(\boldsymbol{a}\times R^\top\boldsymbol{b}) = (R\boldsymbol{a})\times\boldsymbol{b} = \left[R\boldsymbol{a}\right]_\times\boldsymbol{b}, \qquad R\Pi_{\boldsymbol{\theta}}R^\top = R\left(\mathrm{id} - \frac{\boldsymbol{\theta}\boldsymbol{\theta}^\top}{|\boldsymbol{\theta}|^2}\right)R^\top = \Pi_{R\boldsymbol{\theta}}.$$

If we choose $R$ such that $\boldsymbol{\theta} = R\boldsymbol{e}_3$, it follows

$$\begin{aligned}\Pi_{\boldsymbol{\theta}}\left[\boldsymbol{a}\right]_\times\Pi_{\boldsymbol{\theta}} = \Pi_{R\boldsymbol{e}_3}\left[\boldsymbol{a}\right]_\times\Pi_{R\boldsymbol{e}_3} = R\Pi_{\boldsymbol{e}_3}R^\top\left[\boldsymbol{a}\right]_\times R\Pi_{\boldsymbol{e}_3}R^\top &= R(\Pi_{\boldsymbol{e}_3}\left[R^\top\boldsymbol{a}\right]_\times\Pi_{\boldsymbol{e}_3})R^\top\\ &= R((R^\top\boldsymbol{a})\boldsymbol{\cdot}\boldsymbol{e}_3\left[\boldsymbol{e}_3\right]_\times)R^\top = \boldsymbol{a}\boldsymbol{\cdot}\boldsymbol{\theta}\left[\boldsymbol{\theta}\right]_\times\end{aligned}$$

as required. Finally, we can extend this equality pointwise to $\vec{u}$

$$\mathrm{TRT}\left[[\vec{u}]_\times\right](\boldsymbol{\theta},\boldsymbol{x}) = \int_{\mathbb{R}}\Pi_{\boldsymbol{\theta}}\left[\vec{u}(\boldsymbol{x}+t\boldsymbol{\theta})\right]_\times\Pi_{\boldsymbol{\theta}}dt = \int_{\mathbb{R}}\vec{u}(\boldsymbol{x}+t\boldsymbol{\theta})\boldsymbol{\cdot}\boldsymbol{\theta}\left[\boldsymbol{\theta}\right]_\times dt = \mathrm{LRT}[\vec{u}](\boldsymbol{\theta},\boldsymbol{x})\left[\boldsymbol{\theta}\right]_\times.$$

□

*Proof of Theorem 1.4.9.* As the TRT is a linear map, it suffices to prove the theorem in the case $\vec{V} = 0$. Applying the decomposition established in Lemma 1.4.10:

$$\begin{aligned}\mathrm{TRT}\left[\vec{U}\right](\boldsymbol{\theta},\boldsymbol{x}) = 0 &\iff \mathrm{TRT}\left[\mathrm{Sym}(\vec{U})\right](\boldsymbol{\theta},\boldsymbol{x}) + \mathrm{TRT}\left[[\vec{u}]_\times\right](\boldsymbol{\theta},\boldsymbol{x}) = 0 && \text{Lemma 1.4.10(i)}\\ &\iff \mathrm{TRT}\left[\mathrm{Sym}(\vec{U})\right](\boldsymbol{\theta},\boldsymbol{x}) + \mathrm{LRT}[\vec{u}](\boldsymbol{\theta},\boldsymbol{x})\left[\boldsymbol{\theta}\right]_\times = 0 && \text{Lemma 1.4.10(iii)}\\ &\iff \mathrm{TRT}\left[\mathrm{Sym}(\vec{U})\right](\boldsymbol{\theta},\boldsymbol{x}) = \mathrm{Sym}(0) && \text{Lemma 1.4.10(ii)+}\\ &\qquad \text{and } \mathrm{LRT}[\vec{u}](\boldsymbol{\theta},\boldsymbol{x})\left[\boldsymbol{\theta}\right]_\times = 0 - \mathrm{Sym}(0) && \text{sym./skew decomp.}\\ &\iff \mathrm{TRT}\left[\mathrm{Sym}(\vec{U})\right](\boldsymbol{\theta},\boldsymbol{x}) = \mathrm{LRT}[\vec{u}](\boldsymbol{\theta},\boldsymbol{x}) = 0 && \boldsymbol{\theta}\neq\boldsymbol{0}\end{aligned}$$

Hence, by Theorem 1.4.8 we know that $\vec{u}$, and thus the skew component of $\vec{U}$, is never uniquely determined but the symmetric component can be recovered through the equality of Lemma 1.4.10(ii). □

This decomposition, Lemma 1.4.10, is very powerful for understanding the tensor tomography problem from an analytical stand-point. In Theorem 1.4.9, it is used to lift the computations from Sharafutdinov (1994); Desai and Lionheart (2016) to directly compute the kernel of the non-symmetric TRT. On the other hand, results from Desai and Lionheart (2016) also compute the (pseudo) inverses of the LRT and symmetric TRT which could be lifted to an FBP-like pseudo-inverse of the non-symmetric TRT. This in turn allows us to characterise the sensitivity to noise, the inverse problem of non-symmetric tomography is only mildly ill-posed as the singular values decay at only a polynomial rate.

## 1.5 Contributions

The chapters which follow are the main contributions of this thesis. In each chapter we address a different aspect of the mathematics of electron tomography with novel theoretical results and numerical validation.

Chapter 2 proposes a new variational model targeted at reconstruction for limited angle tomography. This model is a modification of the TV reconstruction which attempts to adaptively identify and enhance particular structures during the optimisation. A consequence of this adaptivity is that the problem becomes non-convex and requires the development of a new optimisation algorithm, which is the main analytical contribution of this chapter. The work of this chapter is also published in Tovey et al. (2019). My own contributions cover the analytical and numerical results, and guiding the approach.

Chapter 3 focuses on the development of new mathematical models for electron diffraction which enable the reconstruction of new physical properties from an electron microscope. In particular, we frame the task of strain reconstruction as a tensor tomography inverse problem. The forward model is justified analytically from the first Born approximation and numerically from simulated diffraction patterns. The linearised inverse problem itself is also analysed analytically and numerically to show that strain maps can be recovered accurately from data which is of physically realistic quality. The work of this chapter is also in submission with the Journal of Inverse Problems and a preprint is available Tovey et al. (2020). This project was initiated by collaborators in the Materials Science department in Cambridge who continued to offer their invaluable expertise. The mathematical approach and all of the numerical and analytical results are my own contributions.

Chapter 4 proposes a variant of FISTA which allows for spatially adaptive optimisation in the Banach space setting. The main theoretical contribution of this chapter is the analysis of the proposed algorithm where we prove explicit rates of convergence for problems which could not previously be solved with FISTA. Practically, we observe that the proposed algorithm is faster and uses less memory than the standard FISTA algorithm when trying to approximate minimisers in infinite dimensions. This project was initially formulated by myself and all of the results are my own. This work is in collaboration with Antonin Chambolle who has offered guidance on the scope and avenues are most interesting to pursue.

Chapters 5 and 6 are prospective works which show interesting initial findings but are unfinished. Chapter 5 looks at extending the work of Chapter 4 by designing a basis for more efficient discretisation of TV reconstruction problems. We also demonstrate that current state-of-the-art methods cannot achieve this level of efficiency in general. Chapter 6 presents numerical comparisons of several methods for solving the inverse problem of finding the coordinates of atoms from X-ray data. We identify several factors which can have a large influence on the final reconstruction quality. The work in Chapter 5 was developed in constant discussion with Stephan Hilb during his visit to Cambridge, therefore it is difficult to isolate individual contributions. All of the proofs included here were originally by my own hand. All of the results of Chapter 5 are my own.

A small contribution of this work to the theory of tensor tomography is contained in the preliminaries, namely Theorem 1.4.9 and Lemma 1.4.10. These results will not appear novel to researchers in the field but a concise proof could not be found in the literature. Further clarification is given in the main text.

Another contribution of this PhD has been in the collaboration with more applied researchers. In Collins et al. (2017); Longley et al. (2018); Collins et al. (2019) I contributed the code and inverse problems expertise for electron tomography reconstructions. In Lewis et al. (2020) I advised on the implementation and inversion of a spectral tomography problem. I initiated the work of Tovey and Liang (2020) and provided analytical and numerical assistance throughout.

Finally, in Chapter 7 we summarise the results and possible future work relating to this thesis.

# Chapter 2

# Directional Sinogram Inpainting for Limited Angle Tomography

As was seen in Section 1.4.7, the limited angle tomography inverse problem combines the problems of inpainting, denoising, and indirect reconstruction. In this chapter we propose a new joint model of reconstruction, i.e. where we reconstruct both the original sample and the clean, fully-sampled data.

The inpainting component of this problem is very challenging and so we use a modified form of total variation. It has been shown that imposing a bias on which directions should be penalised more heavily can improve inpainting results. The challenge of this approach is knowing which directions to promote or penalise, the optimal regulariser (choice of directional bias) depends on the optimal reconstruction. As a result of this, the numerical optimisation requires the development of a new optimisation algorithm for functionals which are both non-smooth and non-convex.

We perform numerical experiments on two synthetic datasets and one EM dataset. Our results show consistently that the joint inpainting and reconstruction framework can recover cleaner and more accurate structural information than the current state of the art methods.

## 2.1 Recap of limited angle tomography

The schematic of limited angle data was shown in Figure 1.12 (reproduced in Figure 2.1) and several examples of reconstructions are shown in Figures 1.14 to 1.16. In the context of this chapter, the inverse problem is stated as:

$$\text{given data } \eta\text{, find optimal pair } (u,v) \text{ such that } \mathcal{S}v \approx \eta, \mathcal{R}u \approx v,$$

where $\mathcal{R}$ is the fully-sampled X-ray transform and $\mathcal{S}$ the sampling operator corresponding to the limited angle regime shown in Figure 2.1.

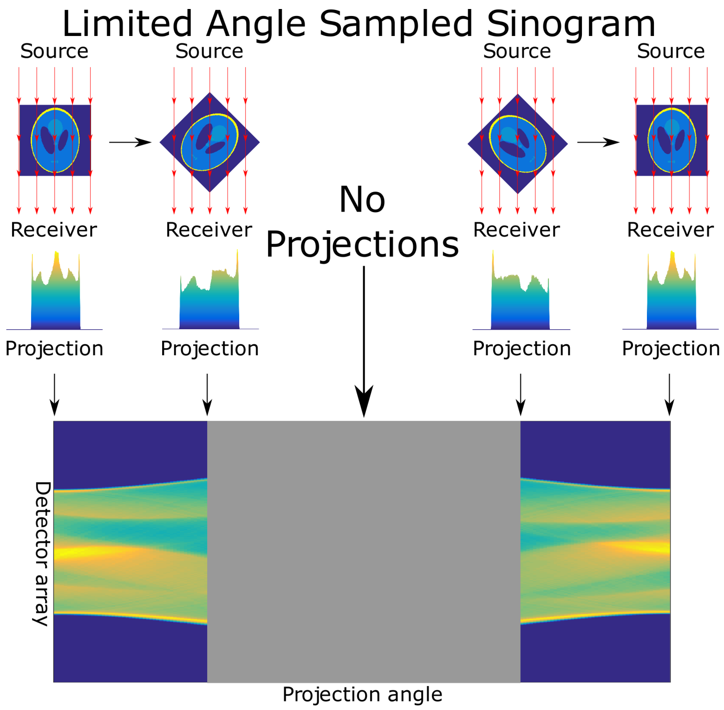


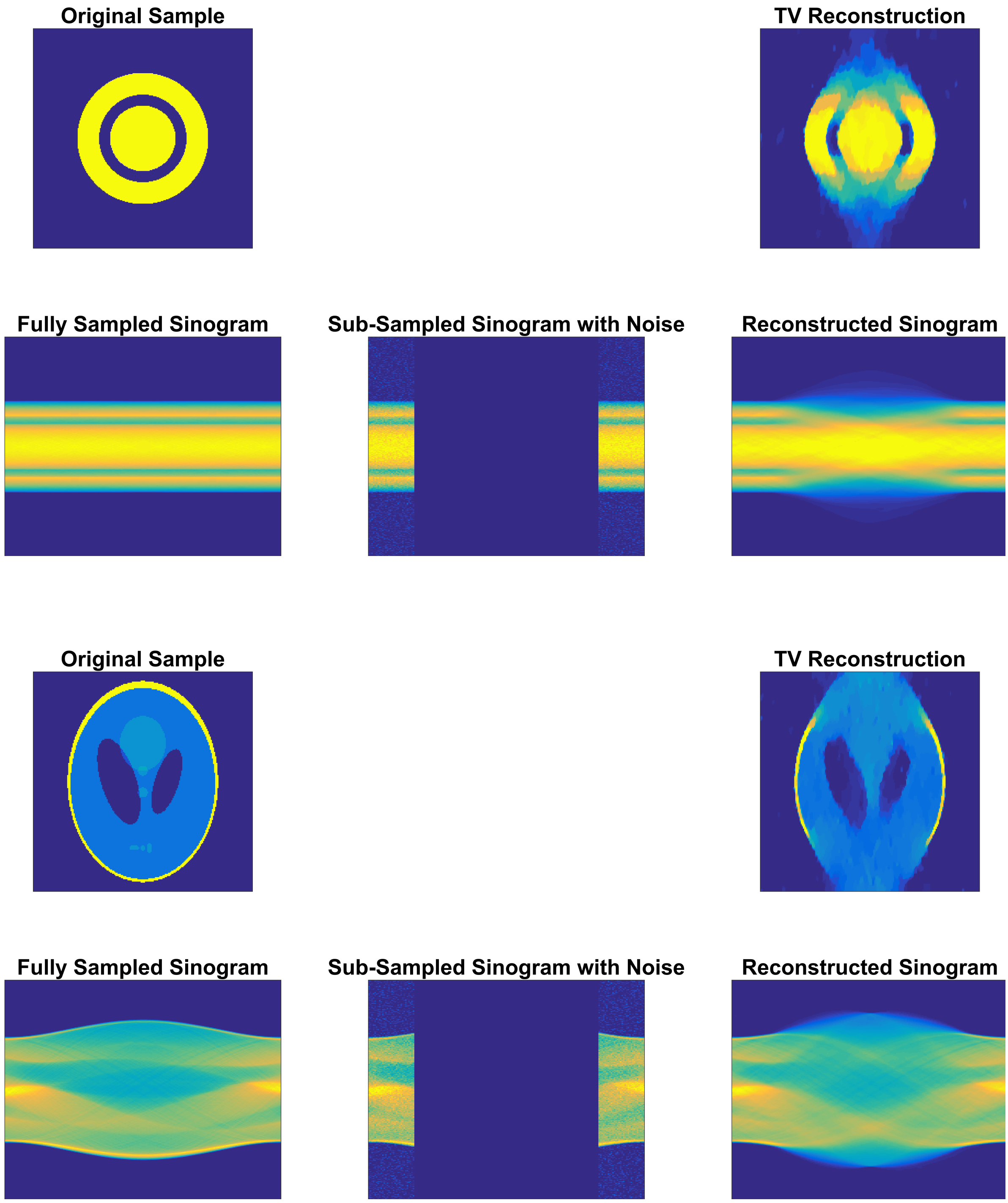
**Figure 2.1** Reproduced from Figure 1.12. Schematic of the acquisition of limited angle acquisition. In the limited angle setting we can only rotate the sample a small amount clockwise and anti-clockwise which results in missing data in the middle of the sinogram.

The most common methodology that has been used to reconstruct pairs $(u, v)$ is to solve each component of the inverse problem sequentially. Typically, we can express the pipeline for such methods as:

$$\begin{aligned} v &= \text{ optimal inpainted sinogram given } \eta, \\ u &= \text{ optimal reconstruction given } v. \end{aligned}$$

This has seen much success in heavy metal artifact reduction (Köstler et al., 2006; Zhang et al., 2011) where a regularisation functional for the inpainting problem may be constructed from dictionary learning (Li et al., 2014), fractional order TV (Zhang et al., 2011), and Euler's Elastica (Gu et al., 2006). Euler's Elastica has also been used in the limited angle problem (Zhang et al., 2017) along with more customised interpolation methods (Kalke and Siltanen, 2014). These approaches use prior knowledge on the sinogram to calculate $v$, and then the spatial prior to calculate $u$ from $v$; at no point is the choice of $v$ influenced by the spatial prior.

Recently, these traditional methods have received a revival through machine learning methods, see for instance Gu and Ye (2017); Hammernik et al. (2017). In both of these examples the main artifact reduction is a learned denoising step which only enforces prior



Original Sample

TV Reconstruction

Fully Sampled Sinogram

Sub-Sampled Sinogram with Noise

Reconstructed Sinogram

Original Sample

TV Reconstruction

Fully Sampled Sinogram

Sub-Sampled Sinogram with Noise

Reconstructed Sinogram

**Figure 2.2** Reproduced from Figure 1.16. Examples when TV reconstructions cannot recover the global structures of samples. When there is a large missing wedge (⅔ of data unseen) and noise on the projections, then reconstructions exhibit characteristic blurring in the vertical direction. This can also be seen in the extrapolated region of the sinograms as a loss of structure.

knowledge on $u$. An interesting alternative was suggested by Bubba et al. (2019) where the inpainting is performed directly on $u$ in a sheerlet basis, again using machine learning.

A full joint approach allows us to use all of our prior knowledge to inform the choice of both $u$ and $v$. If our prior knowledge is consistent with the true data, then this extra utilisation of our prior must have the potential to improve the recovery of both $u$ and $v$. In this chapter we propose a full joint approach which allows us to use all of our prior knowledge at once. To realise this idea, we encode prior knowledge and consistency terms into a single energy functional such that an optimal pair of reconstructions will minimise this joint functional. The joint functional is written as

$$\mathrm{E}(u,v) = \alpha_1 \, \mathrm{D}_1(\mathcal{R}u, v) + \alpha_2 \, \mathrm{D}_2(\mathcal{S}\mathcal{R}u, \eta) + \alpha_3 \, \mathrm{D}_3(\mathcal{S}v, \nu) + \alpha_4 \, \mathrm{G}_1(u) + \alpha_5 \, \mathrm{G}_2(v) \tag{2.1}$$

where $\alpha_i \geq 0$ are weighting parameters, $\mathrm{D}_i$ are appropriate distance functionals, and $\mathrm{G}_i$ are regularisation functionals which encode our prior knowledge. Note that choice of $\mathrm{D}_2$ and $\mathrm{D}_3$ are dictated by the data noise model. In what follows, $\mathrm{G}_1$ is chosen to be the total variation.

Our choice for $\mathrm{G}_2$, the sinogram regularisation, is based on theoretical and heuristic observations. Thirion (1991) has shown that discontinuities in $u$ correspond to sharp edges in the sinogram. In Figure 2.2, we also see that blurred reconstructions correspond to loss of structure in the sinogram. Therefore, $\mathrm{G}_2$ will be chosen to detect sharp features in the given data and preserve these through the inpainting region. The exact form of $\mathrm{G}_2$ will be formalised in Section 2.3.

A typical advantage of joint models is that they generalise sequential ones. For instance, if we let $\alpha_2, \alpha_4 \to \infty$, then we recover the TV reconstruction model. Alternatively, if we let $\alpha_3, \alpha_5 \to \infty$, then we recover a method which performs the inpainting and then the reconstruction sequentially, as in Zhang et al. (2011); Gu et al. (2006); Zhang et al. (2017). Recent work by Burger et al. (2014) has shown that such a joint approach can be advantageous in similar inverse problems, but closest to our approach is that of Dong et al. (2013) where $\mathrm{G}_1$ and $\mathrm{G}_2$ were chosen to encode wavelet sparsity in both $u$ and $v$. We shall demonstrate that the flexibility of our joint model, (2.1), can allow for a better adaptive fitting to the data.

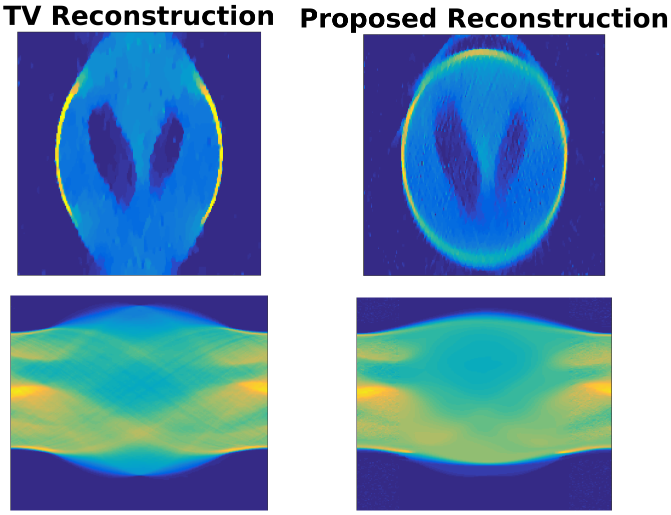


**Figure 2.3** Demonstration of the improvement which can be achieved by using a model as in (2.1). Left hand images show state of the art reconstructions using total variational regularisation ($\alpha_1 = \alpha_3 = \alpha_5 = 0$). This reconstruction clearly shows the characteristic blurring artifacts at the top and bottom. In our proposed joint reconstruction (right hand) we can minimise these effects.

### 2.1.1 Overview and contributions

The main contribution of this work is to provide a framework for building models of the form described in (2.1) and provide new proofs for a numerical scheme for minimising these functionals. This numerical scheme is valid for a very large class of non-smooth and non-convex functionals $\mathrm{G}_i$ and thus could be used in many other applications.

Section 2.2 first outlines the necessary concepts and notation needed to formalise the statement of our specific joint model in Section 2.3. It will become apparent that the main numerical requirements of this work will require minimising a functional which is neither convex or smooth. Section 2.4 will start by reviewing recent work by Ochs et al. (2019), and we then provide alternative concise and self-contained proofs. Our main contribution here is to extend the existing results to an alternating (block) descent scenario. Finally, we present numerical results including two synthetic phantoms and experimental electron microscopy data where the limited angle situation arises naturally.

## 2.2 Preliminaries

### 2.2.1 Directional total variation regularisation

For our sinogram regularisation functional we shall use a directionally weighted TV penalty, motivated by the TV diffusion model developed by Weickert (1998) for various imaging techniques including denoising, inpainting and compression. Such an approach has already shown great ability for enhancing edges in noisy or blurred images, and preserves line structures across inpainting regions (Berkels et al., 2006; Estellers et al., 2015; Bertalmio et al., 2000). The heuristic for our regularisation on the sinogram was described in Figure 2.2 and we shall encode it in an anisotropic TV penalty which shall be written as

$$\mathrm{DTV}(v) = \int |\mathcal{B}(\boldsymbol{x})\nabla v(\boldsymbol{x})| d\boldsymbol{x} = \|\mathcal{B}\nabla v\|_{1,2} \text{ for some anisotropic } \mathcal{B}\colon \mathbb{R}^2 \to \mathbb{R}^{2\times 2}.$$

The power of such a weighted extension of TV is that once a line is detected, either known beforehand or detected adaptively, we can embed this knowledge in $\mathcal{B}$ and enhance or sharpen that structure in the final result. The general form that we choose for $\mathcal{B}$ is

$$\begin{aligned}\mathcal{B}(\boldsymbol{x}) = {} & c_1(\boldsymbol{x})\boldsymbol{e}_1(\boldsymbol{x})\boldsymbol{e}_1(\boldsymbol{x})^\top + c_2(\boldsymbol{x})\boldsymbol{e}_2(\boldsymbol{x})\boldsymbol{e}_2(\boldsymbol{x})^\top \\ & \text{such that } \boldsymbol{e}_i\colon \mathbb{R}^2 \to \mathbb{R}^2, |\boldsymbol{e}_i(\boldsymbol{x})| = 1, \boldsymbol{e}_1(\boldsymbol{x})\cdot\boldsymbol{e}_2(\boldsymbol{x}) = 0, \end{aligned} \tag{2.2}$$

in other words

$$\mathrm{DTV}(v) = \int \sqrt{c_1^2|\boldsymbol{e}_1\cdot\nabla v|^2 + c_2^2|\boldsymbol{e}_2\cdot\nabla v|^2} d\boldsymbol{x}.$$

Examples of this are presented in Figure 2.4. Note that the choice $c_1 = c_2$ recovers the traditional TV regulariser but, when $|c_1| \ll c_2$, we allow for much larger (sparse) gradients in the direction of $\boldsymbol{e}_1$. This allows for large jumps in the direction of $\boldsymbol{e}_1$, whilst maintaining flatness in the direction of $\boldsymbol{e}_2$. We use a regularisation framework similar to that proposed by Kaipio et al. (1999), however our choice of parametrisation more closely follows that of Weickert (1998). Given a noisy image, $\nu$, we can construct the structure tensor

$$(\nabla\nu_\rho\nabla\nu_\rho^\top)_\sigma(\boldsymbol{x}) = \lambda_1(\boldsymbol{x})\boldsymbol{e}_1(\boldsymbol{x})\boldsymbol{e}_1(\boldsymbol{x})^\top + \lambda_2(\boldsymbol{x})\boldsymbol{e}_2(\boldsymbol{x})\boldsymbol{e}_2(\boldsymbol{x})^\top \text{ such that } \lambda_1(\boldsymbol{x}) \geq \lambda_2(\boldsymbol{x}) \geq 0$$

where

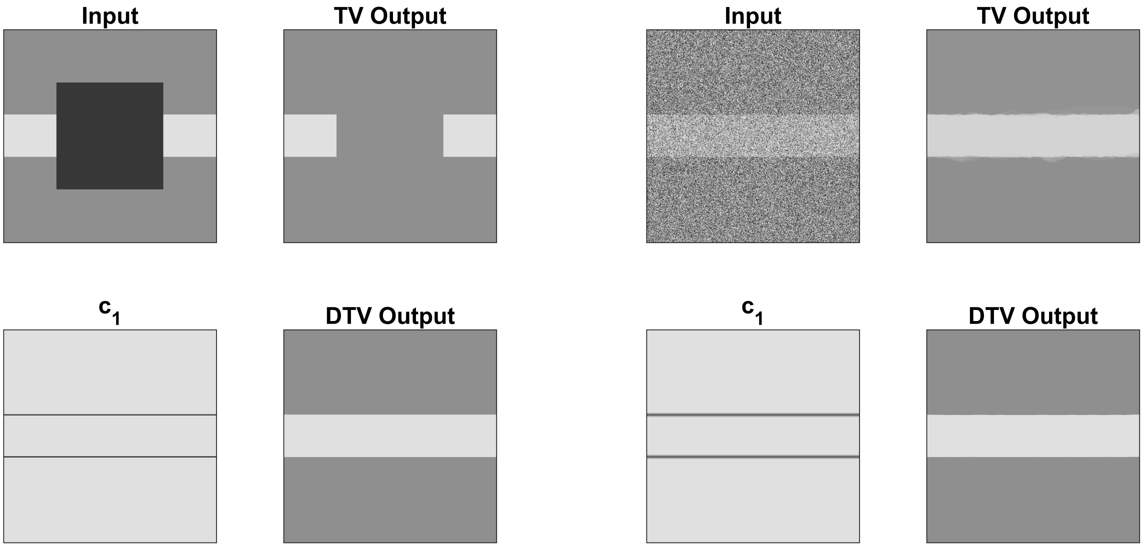
$$\nu_\rho(\boldsymbol{x}) = \left[\nu \star \exp\left(-\tfrac{|\cdot|^2}{2\rho^2}\right)\right](\boldsymbol{x}) = \int \exp\left(-\tfrac{|\boldsymbol{y}-\boldsymbol{x}|^2}{2\rho^2}\right)\nu(\boldsymbol{y})$$

denotes convolution with the heat kernel, the same notation is used for subscript $\sigma$. This eigenvalue decomposition is typically very informative in constructing $\mathcal{B}$. If we define

$$\Delta(\boldsymbol{x}) = \lambda_1(\boldsymbol{x}) - \lambda_2(\boldsymbol{x}) \text{ as coherence,} \qquad \Sigma(\boldsymbol{x}) = \lambda_1(\boldsymbol{x}) + \lambda_2(\boldsymbol{x}) \text{ as energy,}$$

then the eigenvectors give the alignment of edges and $\Delta\backslash\Sigma$ characterise the local behaviour, as in Figure 2.5. In particular, we simplify the model to

$$\mathcal{B}_\nu(\boldsymbol{x}) \coloneqq c_1(\boldsymbol{x}|\Delta,\Sigma)\boldsymbol{e}_1(\boldsymbol{x})\boldsymbol{e}_1(\boldsymbol{x})^\top + c_2(\boldsymbol{x}|\Delta,\Sigma)\boldsymbol{e}_2(\boldsymbol{x})\boldsymbol{e}_2(\boldsymbol{x})^\top \tag{2.3}$$



**(a)** Inpainting without noise

**(b)** Denoising

**Figure 2.4** Examples comparing TV with directional TV for inpainting and denoising. Both examples have the same edge structure and so parameters in (2.2) are the same in both. DTV uses $c_2 = 1$ and $c_1$ as the indicator (0 or 1) shown in the bottom left plot, TV is the case $c_1 = c_2 = 1$. Panel (a): Top left image is inpainting input where the dark square shows the inpainting region. The structure of $c_1$ allows DTV (bottom right) to connect lines over arbitrary distances, whereas TV inpainting (top right) will fail to connect the lines if the horizontal distance is greater than the vertical separation of the edges. Panel (b): Top left image is denoising input. DTV has two advantages. Firstly, the structure of $c_1$ recovers a much straighter line than that in the TV reconstruction. Secondly, TV penalises jumps equally in each direction and so the contrast is reduced, DTV is able to penalise noise oscillations independently from edge discontinuities which allows us to maintain much higher contrast.

where the only parameters left to choose are $c_i$. Typical examples of include

$$c_1 = \frac{1}{\sqrt{1+\Sigma^2}}, \qquad c_2 = 1,$$

and

$$c_1 = \varepsilon, \qquad c_2 = \varepsilon + \exp\left(-{}^{1}\!/\!{}_{\Delta^2}\right) \text{ for some } \varepsilon > 0.$$

The key idea here is that $c_1 \ll c_2$ near edges and $c_1 = c_2$ on flat regions. In practice, $\nu$ will also be an optimisation parameter and so we shall require a regularity result on our choice of $\nu \mapsto \mathcal{B}_\nu$, now characterised uniquely by our choice of $c_i$.

**Theorem 2.2.1.** *If*

1. *$c_i$ are $2k$ times continuously differentiable in $\Delta$ and $\Sigma$, $k \geq 1$,*

2. *$c_1(\boldsymbol{x}|0,\Sigma) = c_2(\boldsymbol{x}|0,\Sigma)$ for all $\boldsymbol{x}$ and $\Sigma \geq 0$,*

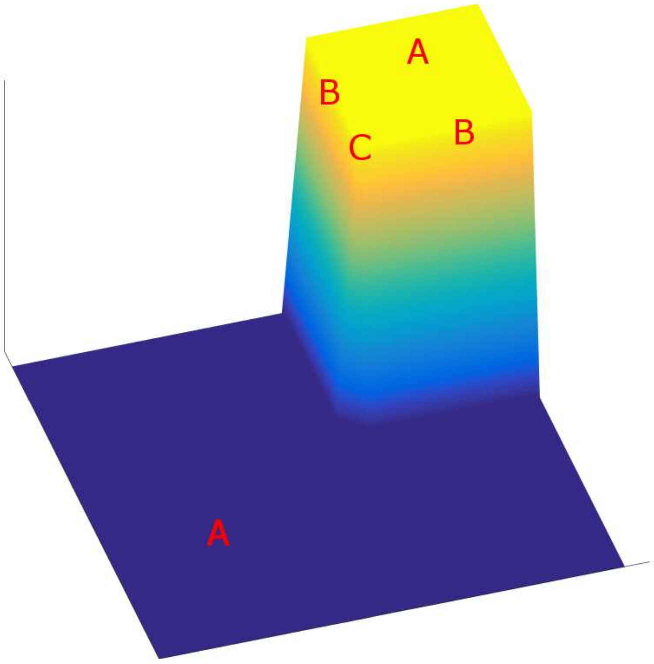


**Figure 2.5** Surface representing a characteristic image, $\nu$, to demonstrate the behaviour of $\Sigma$ and $\Delta$. Away from edges (A) we have $\Sigma \approx \Delta \approx 0$. On simple edges (B) we have $\Sigma \approx \Delta \gg 0$ and, finally, at corners (C) we have $\Sigma \gg \Delta$.

*3. and* $\nabla_\Delta^{2j-1} c_1(\boldsymbol{x}|0,\Sigma) = \nabla_\Delta^{2j-1} c_2(\boldsymbol{x}|0,\Sigma) = 0$ *for all* $\boldsymbol{x}$ *and* $\Sigma \geq 0, j = 1\ldots,k$,

*then* $\mathcal{B}_\nu$ *is* $C^{2k-1}$ *with respect to* $\nu$ *for all* $\rho > 0, \sigma \geq 0$.

The proof of Theorem 2.2.1 is contained in Section A.1.

**Remark 2.2.2.** *In the proof of Theorem 2.2.1:*

- *Property (ii) is necessary for* $\mathcal{B}_\nu$ *to be well defined and continuous for all fixed* $\nu$
- *If we can write* $c_i = c_i(\Delta^2, \Sigma)$, *then property (iii) holds trivially*

Particular to the context of this work, $\nu$ will be periodic. Sinograms are naturally $2\pi$-periodic with respect to the angle parameter $x_1$, and we will only consider examples $u^\dagger$ which are compactly supported in space therefore $\nu$ is compactly supported with respect to the detector location $x_2$. This ensures that the convolutions with Gaussian kernels can be computed with Fourier methods without any negative boundary effects.

## 2.3 The joint model

Now that all of the notation and concepts have been defined, we can formalise the statement of our particular joint model of the form in (2.1). The variables in question are:

- $\mathcal{R}\colon L^1(\mathbb{R}^2) \to L^1(T\mathbb{S}^1)$, is the fully sampled X-ray transform
- $\mathcal{S}\colon L^1(T\mathbb{S}^1) \to L^1(\Omega')$ is the sampling operator on some $\Omega' \Subset \mathbb{S}^1 \times \mathbb{R}$
- The desired reconstructed sample is $u \in \mathbb{BV}(\Omega, \mathbb{R})$ on some domain $\Omega$
- The noisy sub-sampled data is $\eta \in L^1(\Omega')$ which we extend such that $\eta|_{\Omega'^c} = 0$ for notational convenience.
- The full reconstructed sinogram is $v \in L^1(\mathbb{S}^1 \times \mathbb{R})$

We combine this with our choice of data fidelities and prior functions into the model

$$(u, v) = \operatorname*{argmin}_{u\geq 0} \mathrm{E}(u, v) = \operatorname*{argmin}_{u\geq 0} \frac{1}{2}\left\|\mathcal{R}u - v\right\|_{\alpha_1}^2 + \frac{\alpha_2}{2}\left\|\mathcal{S}\mathcal{R}u - \eta\right\|_2^2 + \frac{\alpha_3}{2}\left\|\mathcal{S}v - \eta\right\|_2^2 + \beta_1 \operatorname{TV}(u) + \beta_2 \operatorname{DTV}_{\mathcal{R}u}(v) \quad (2.4)$$

where

$$\operatorname{DTV}_{\mathcal{R}u}(v) = \left\|\mathcal{B}_{\mathcal{R}u}\nabla v\right\|_{1,2}$$

and $\alpha_i, \beta_i > 0$ are weighting parameters, $\mathcal{B}_{\mathcal{R}u}$ as defined in (2.3). The value $\alpha_1$ is embedded in the norm because it is a spatially varying weight, taking different (constant) values inside and outside of $\Omega'$. We note that the norms involving $\eta$ are determined by the noise model of the acquisition process, in this case Gaussian noise. The final metric pairing $\mathcal{R}u$ and $v$ was free to be chosen to promote any structural similarity between the two quantities. We have chosen the squared $L^2$ norm for simplicity, although if some structure is known to be important, then there is a wide choice of specialised functions from which to choose (c.f. Ehrhardt et al., 2015).

The choice of regularisation functionals reflects prior assumptions on the expected type of sample; all of the examples shown later will follow these assumptions. The isotropic TV penalty is chosen as $u$ is expected to be piecewise constant. This will reduce oscillations from $u$ and favour 'stair-case'-like changes of intensity over smooth ones. The assumptions of our regularisation on $v$ must also be derived from expected properties of $u$. What is known from Thirion (1991), and can be seen in Figure 2.2, is that discontinuities of $u$ along curves in the spatial domain, say $\gamma$, generate a so called *dual curve* in the sinogram. $\mathcal{R}u$ will also have an edge, although possibly not a discontinuity, along this dual curve. Thus, perpendicular to the dual curve $v$ should have sharp features, and parallel to the dual curve intensity should vary slowly. The assumption of our regularisation is that, if a dual curve is present in the visible component of the data, then it should correspond to some $\gamma$ in the spatial domain. The extrapolation of this dual curve must behave like the boundary of a level set of $u$, i.e. preserve the sharp edge and slow varying intensities in $v$. The main influence of this regularisation is in the inpainting region, and so any artifacts it introduces should also only affect edges corresponding to these invisible singularities, including streaking artifacts. Another bias that is present is an assumption that dual curves are themselves smooth. In the inpainting region, this will encourage dual curves with low curvature thus invisible singularities are likely to follow near-circular arcs in the spatial domain. Final parameter choices, such as $\alpha_i, \beta_i$ and $c_i$, are not necessary at this point and will be chosen in Section 2.5.1.

The immediate question to ask is whether this model (2.4) is well posed. For a non-convex function we typically cannot expect to find global minimisers numerically, but the following result shows we can expect some convergence to local minima. Theorem 2.3.1 justifies looking for minima of (2.4).

**Theorem 2.3.1.** *If*

- *$c_i$ are bounded away from 0,*
- *$\rho > 0$,*
- *and $\mathcal{B}_\nu$ is differentiable in $\nu$,*

*then sublevel sets of* $\mathrm{E}$ *are weakly compact in $L^2(\Omega) \times L^2(\mathbb{R}^2)$ and* $\mathrm{E}$ *is weakly lower semi-continuous. i.e. for all $(u_n, v_n) \in L^2(\Omega) \times L^2(\mathbb{R}^2)$*

$$\textit{if} \quad \{\mathrm{E}(u_n, v_n) \text{ s.t. } n \in \mathbb{N}\} \textit{ is uniformly bounded, then a subsequence converges weakly,}$$

$$\liminf_{n\to\infty} \mathrm{E}(u_n, v_n) \geq \mathrm{E}(u, v) \textit{ whenever } u_n \rightharpoonup u, v_n \rightharpoonup v.$$

The proof of this theorem is contained in Section A.2. This theorem justifies numerical minimisation of $\mathrm{E}$ because it tells us that all descending sequences $(\mathrm{E}(u_n, v_n) \leq \mathrm{E}(u_{n-1}, v_{n-1}))$ have a convergent subsequence and any limit point must minimise $\mathrm{E}$ over the original sequence.

## 2.4 Numerical solution

To address the issue of convergence, we shall first generalise our functional and prove the result in the general setting. Functionals will be of the form $\mathrm{E}\colon \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ where $\mathbb{U}$ and $\mathbb{V}$ are Banach spaces and $\mathrm{E}$ accepts the decomposition

$$\mathrm{E}(u, v) = \mathrm{f}(u, v) + \mathrm{g}(J(u, v))$$

such that:

- Sublevel sets of $\mathrm{E}$ are weakly compact. (2.5)
- $\mathrm{f}\colon \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ is jointly convex in $(u, v)$ and bounded from below. (2.6)
- $\mathrm{g}\colon \mathbb{W} \to \mathbb{R}$ is a semi-norm on Banach space $\mathbb{W}$, i.e. for all $t \in \mathbb{R}$, $w, w_1, w_2 \in \mathbb{W}$

$$\mathrm{g}(w) \geq 0, \quad \mathrm{g}(tw) = |t|\,\mathrm{g}(w) \quad \text{and} \quad \mathrm{g}(w_1 + w_2) \leq \mathrm{g}(w_1) + \mathrm{g}(w_2). \tag{2.7}$$

- $J\colon \mathbb{U}\times\mathbb{V}\to\mathbb{W}$ is $C^1$ and for all $K \Subset \mathbb{U}\times\mathbb{V}$, there exists $L_u, L_v < \infty$ such that

$$\mathrm{g}(J(u+du,v) - J(u,v) - \nabla_u J(u,v)du) \leq L_u \left\|du\right\|_{\mathbb{U}}^2 \tag{2.8}$$

$$\mathrm{g}(J(u,v+dv) - J(u,v) - \nabla_v J(u,v)dv) \leq L_v \left\|dv\right\|_{\mathbb{V}}^2 \tag{2.9}$$

  for all $(u,v)\in K$.

Note that if g is a norm, then $L_u$ can be chosen to be the standard Lipschitz factor of $\nabla_u J$. If $J$ is twice Frèchet-differentiable, then these constants must be finite. In our case:

$$\begin{aligned}
\mathrm{f}(u,v) &= \frac{1}{2}\left\|\mathcal{R}u - v\right\|_{\alpha_1}^2 + \frac{\alpha_2}{2}\left\|\mathcal{S}\mathcal{R}u - \eta\right\|^2 \\
&\qquad + \frac{\alpha_3}{2}\left\|\mathcal{S}v - \eta\right\|^2 + \beta_1 \,\mathrm{TV}(u) + \begin{cases} 0 & u \geq 0 \\ \infty & \text{else}\end{cases} \\
\mathrm{g}(w) &= \beta_2 \left\|w\right\|_{2,1} \\
J(u,v) &= \mathcal{B}_{\mathcal{R}u}\nabla v \implies \tau_u \sim \beta_2 \left\|\nabla^2\mathcal{B}.\right\| \left\|\mathcal{R}\right\| \mathrm{TV}(v),\ \tau_v = 0
\end{aligned}$$

Finiteness of $\left\|\nabla^2\mathcal{B}\right\|$ and weak compactness of sublevel sets are given by Theorems 2.2.1 and 2.3.1 respectively. The alternating descent algorithm is stated in Algorithm 2.1. Classical alternating proximal descent would define $u_{n+1} = \operatorname{argmin} \mathrm{E}(u,v_n) + \tau_u \left\|u-u_n\right\|_2^2$. However, because of the complexity of $A_{\mathcal{R}u}$, each sub-problem would have the same complexity as the full problem, making it computationally infeasible. By linearising $\mathcal{B}_\nu$ we overcome this problem as both sub-problems are convex and a standard solver such as Chambolle and Pock (2011); Mosek ApS (2010) may be employed. This second approach is similar to that of the ProxDescent algorithm (Drusvyatskiy and Lewis, 2018; Ochs et al., 2019). We extend the ProxDescent algorithm to cover alternating descent and achieve equivalent convergence guarantees. Using Algorithm 2.1, our statement of convergence is the following theorem.

**Algorithm 2.1**

```
  Input  any $u_0 \in \mathbb{U}$, $\tau_u, \tau_v \geq 0$.
  $n \leftarrow 0$
  while not converged do
      $n \leftarrow n+1$
    $u_n = \operatorname*{argmin}_{u\in\mathbb{U}} \mathrm{f}(u,v_{n-1}) + \tau_u \left\|u-u_{n-1}\right\|_{\mathbb{U}}^2 + \mathrm{g}(J(u_{n-1},v_{n-1}) + \nabla_u J(u_{n-1},v_{n-1})(u-u_{n-1}))$
                                                                                    (2.10)
    $v_n = \operatorname*{argmin}_{v\in\mathbb{V}} \mathrm{f}(u_n,v) + \tau_v \left\|v-v_{n-1}\right\|_{\mathbb{V}}^2 + \mathrm{g}(J(u_n,v_{n-1}) + \nabla_v J(u_n,v_{n-1})(v-v_{n-1}))$   (2.11)
  end while
  return  $(u_n, v_n)$
```

**Theorem 2.4.1** (Convergence of alternating minimisation)**.** *If* $\mathrm{E}$ *satisfies* (2.5)-(2.9) *and* $(u_n, v_n)$ *are a sequence generated by Algorithm 2.1, then*

- $\mathrm{E}(u_{n+1}, v_{n+1}) \leq \mathrm{E}(u_n, v_n)$ *for each* $n$*.*
- *A subsequence of* $(u_n, v_n)$ *must converge weakly in* $\mathbb{U} \times \mathbb{V}$
- *If* $\{(u_n, v_n) \text{ s.t. } n =\in \mathbb{N}\}$ *is contained in a finite dimensional space, then every limit point of* $(u_n, v_n)$ *must be a critical point (as will be defined in Definition 2.4.4) of* $\mathrm{E}$ *in both the direction of* $u$ *and* $v$*.*

This result is the parallel of Lemma 10, Theorem 18, and Corollary 21 in Ochs et al. (2019) without the alternating or block descent setting. There is some overlap in the analysis for the two settings, although we present an independent proof which is more direct and we feel gives more intuition for our more restricted class of functionals. The rest of this section is now dedicated to the proof of Theorem 2.4.1.

For notational convenience we shall compress notation such that:

$$\mathrm{f}_{n,m} = \mathrm{f}(u_n, v_m), \quad J_{n,m} = J(u_n, v_m), \quad \mathrm{E}_{n,m} = \mathrm{E}(u_n, v_m).$$

### 2.4.1 Sketch proof

The proof of Theorem 2.4.1 will be the consequence of two lemmas.

- Let $\tau_u \backslash \tau_v$ be as defined in Algorithm 2.1. In Lemma 2.4.3 we show for $\tau_u, \tau_v$ sufficiently large, the sequence $\mathrm{E}_{n,n}$ is monotonically decreasing and sequences $\|u_n - u_{n-1}\|_{\mathbb{U}}$, $\|v_n - v_{n-1}\|_{\mathbb{V}}$ converge to 0. This follows by a relatively standard sufficient decrease argument as seen in Pock and Sabach (2016); Ochs et al. (2019); Liang et al. (2016).
- In Definition 2.4.4 we define the notion of a critical point for functions which are non-convex and non-differentiable. This follows the work of Drusvyatskiy et al. (2019).
- In Lemma 2.4.6 we show that any convergent sequence of Algorithm 2.1 must be a critical point in $u$ and a critical point in $v$. Note that it is very difficult to expect more than this in such a general setting, for instance Example 2.4.2 shows a uniformly convex energy where this statement is sharp. The common technique for overcoming this is assuming differentiability in the terms including both $u$ and $v$ (Pock and Sabach, 2016; Ochs et al., 2014; Bolte et al., 2014). These previous results and algorithms are not available to us as we allow non-convex terms which are also non-differentiable.

**Example 2.4.2.** *Define* $\mathrm{E}(u, v) = \max(u, v) + u^2 + v^2$ *for* $u, v \in \mathbb{R}$*. This is clearly jointly convex in* $(u, v)$ *and thus a simple case of functions considered in Theorem 2.4.1. Suppose*

$(u_0, v_0) = (0, 0)$, *then*

$$\begin{aligned} u_1 &= \operatorname{argmin} \mathrm{E}(u, v_0) + \tau_u (u - u_0)^2 = 0 \\ v_1 &= \operatorname{argmin} \mathrm{E}(u_1, v) + \tau_v (v - v_0)^2 = 0 \end{aligned}$$

*Hence* $(0, 0)$ *is a limit point of the algorithm but it is not a critical point,* $\mathrm{E}$ *is strongly convex and so it has only one critical point,* $(-{}^1\!/\!_2, -{}^1\!/\!_2)$.

### 2.4.2 Sufficient decrease lemma

In the following we prove the monotone decrease property of our energy functional between iterations.

**Lemma 2.4.3.** *If*

$$\tau_u \geq L_u + \tau_U, \qquad \tau_v \geq L_v + \tau_V$$

*for some* $\tau_U, \tau_V \geq 0$, *then*

$$\sum^{\infty} \tau_U \|u_n - u_{n-1}\|_{\mathbb{U}}^2 + \tau_V \|v_n - v_{n-1}\|_{\mathbb{V}}^2 \leq \mathrm{E}(u_0, v_0)$$

*and*

$$\mathrm{E}(u_{n+1}, v_{n+1}) \leq \mathrm{E}(u_n, v_n) \textit{ for all } n.$$

*Proof.* Note by Equations (2.10) and (2.11) (definition of $u_n \backslash v_n$), we have

$$\mathrm{f}_{n+1,n} + \mathrm{g}(J_{n,n} + \nabla_u J_{n,n}(u_{n+1} - u_n)) + \tau_u \|u_{n+1} - u_n\|_{\mathbb{U}}^2 \leq \mathrm{E}_{n,n} \tag{2.12}$$

$$\mathrm{f}_{n+1,n+1} + \mathrm{g}(J_{n+1,n} + \nabla_v J_{n+1,n}(v_{n+1} - v_n)) + \tau_v \|v_{n+1} - v_n\|_{\mathbb{V}}^2 \leq \mathrm{E}_{n+1,n} \tag{2.13}$$

Further, by application of the triangle inequality for g and the mean value theorem we have

$$\begin{aligned} \mathrm{g}(J_{n+1,n}) - &\mathrm{g}(J_{n,n} + \nabla_u J_{n,n}(u_{n+1} - u_n)) + \tau_U \|u_{n+1} - u_n\|_{\mathbb{U}}^2 \\ &\leq \mathrm{g}(J_{n+1,n} - J_{n,n} - \nabla_u J_{n,n}(u_{n+1} - u_n)) + \tau_U \|u_{n+1} - u_n\|_{\mathbb{U}}^2 \\ &= \mathrm{g}([\nabla_u J(\widetilde{u}) - \nabla_u J_{n,n}](u_{n+1} - u_n)) + \tau_U \|u_{n+1} - u_n\|_{\mathbb{U}}^2 \\ &\leq \mathrm{Lip}_{\mathbb{U},\mathrm{g}}(\nabla_u J(\cdot, v_n)) \|u_{n+1} - u_n\|_{\mathbb{U}}^2 + \tau_U \|u_{n+1} - u_n\|_{\mathbb{U}}^2 \\ &\leq \tau_u \|u_{n+1} - u_n\|_{\mathbb{U}}^2 \end{aligned} \tag{2.14}$$

By equivalent argument,

$$\mathrm{g}(J_{n+1,n+1}) - \mathrm{g}(J_{n,n+1} + \nabla_v J_{n+1,n}(v_{n+1} - v_n)) + \tau_V \|v_{n+1} - v_n\|_{\mathbb{V}}^2 \leq \tau_v \|v_{n+1} - v_n\|_{\mathbb{V}}^2 \tag{2.15}$$

Summing Equations (2.12) to (2.15) gives

$$\mathrm{E}_{n+1,n+1} + \tau_U \left\|u_{n+1} - u_n\right\|_{\mathbb{U}}^2 + \tau_V \left\|v_{n+1} - v_n\right\|_{\mathbb{V}}^2 \leq \mathrm{E}_{n,n}$$

This immediately gives the monotone decrease property of $\mathrm{E}_{n,n}$. If we also sum this over $n$, then we achieve the final statement of the theorem:

$$\sum_{n=1}^{\infty} \tau_U \left\|u_{n+1} - u_n\right\|_{\mathbb{U}}^2 + \tau_V \left\|v_{n+1} - v_n\right\|_{\mathbb{V}}^2 \leq \mathrm{E}_{0,0} - \lim \mathrm{E}_{n,n} \leq \mathrm{E}_{0,0}\,.$$

□

### 2.4.3 Convergence to critical points

First, following the work of Drusvyatskiy et al. (2019), we define criticality in terms of the slope of a function.

**Definition 2.4.4.** *We shall say that $u^*$ is a critical point of* $\mathrm{F}\colon \mathbb{U} \to \mathbb{R}$ *if*

$$|\partial\, \mathrm{F}(u^*)| = 0$$

*where we define* the slope *of* F *at* $u^*$ *to be*

$$|\partial\, \mathrm{F}(u^*)| = \limsup_{du \to 0} \frac{\max(0, \mathrm{F}(u^*) - \mathrm{F}(u^* + du))}{\|du\|}$$

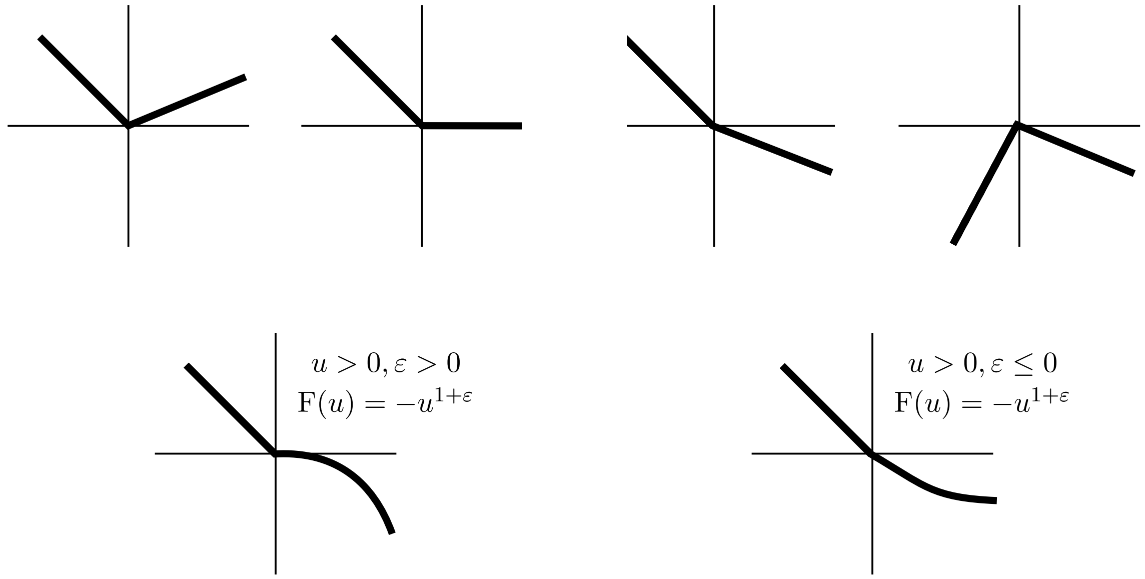
The first point to note is that this definition generalises the concept of a first order critical point for both smooth functions and convex functions (in terms of the convex sub-differential). In particular, if $\mathrm{F} \in C^1$, then

$$\begin{gathered} |\partial\, \mathrm{F}(u^*)| = \max\left(0, \sup_{\|du\|=1} -\langle \nabla \mathrm{F}(u^*),\ du\rangle\right) = \|\nabla \mathrm{F}(u^*)\|\,, \\ \text{hence, } |\partial\, \mathrm{F}(u^*)| = 0 \iff \|\nabla \mathrm{F}(u^*)\| = 0 \iff \nabla \mathrm{F}(u^*) = 0. \end{gathered}$$

Similarly, if F is convex, then

$$\begin{gathered} u^* \in \operatorname{argmin} F \iff \mathrm{F}(u^*) \leq \mathrm{F}(u^* + du) \text{ for all } du, \\ \text{hence, } |\partial\, \mathrm{F}(u^*)| = 0 \iff \forall du, 0 \geq \frac{\mathrm{F}(u^*) - \mathrm{F}(u^* + du)}{\|du\|} \iff u^* \in \operatorname{argmin} \mathrm{F}\,. \end{gathered}$$

For a differentiable function we cannot tell whether a critical point is a local minimum, maximum or saddle point. In general, this is also true for Definition 2.4.4, however, at points of non-differentiability there is a bias towards local minima. This can be seen in the following example.


$u > 0, \varepsilon > 0$
$\mathrm{F}(u) = -u^{1+\varepsilon}$
$u > 0, \varepsilon \leq 0$
$\mathrm{F}(u) = -u^{1+\varepsilon}$

**(a)** Examples of Critical Points

**(b)** Examples of Non-Critical Points

**Figure 2.6** Examples of 1D functions where 0 is/isn't a critical point by Definition 2.4.4. If a function is piecewise linear, then 0 is a critical point iff each directional derivative is non-negative. If the function can be approximated on an interval of $u > 0$ to first order terms by $\mathrm{F}(u) = cu^{1+\varepsilon}$, then criticality can be characterised sharply. If $c \geq 0$, then 0 is always a critical point. If $c < 0$, then 0 is a critical point iff $\varepsilon > 0$, however, 0 is never a local minimum.

**Example 2.4.5.** *Consider* $\mathrm{F}(u) = -\|u\|$, *then*

$$|\partial \mathrm{F}(0)| = \limsup_{du\to 0} \max\left(0, \frac{0 + \|0 + du\|}{\|du\|}\right) = 1.$$

*Hence, 0 is not a critical point of* $\mathrm{F}$. *This bias is due to the* lim sup *in the definition which detects the strictly negative directional derivatives. This doesn't affect smooth functions as directional derivatives must vanish continuously to 0 about a critical point.*

Some more examples are shown in Figure 2.6. Now we shall show that our iterative sequence converges to a critical point in this sense.

**Lemma 2.4.6.** *If* $(u_n, v_n)$ *are as given by Algorithm 2.1 and* $\mathbb{U}, \mathbb{V}$ *are finite dimensional spaces, then every limit point of* $(u_n, v_n)$, *e.g.* $(u^*, v^*)$, *is a critical point of* $\mathrm{E}$ *in each coordinate direction. i.e.*

$$|\partial_u \mathrm{E}(u^*, v^*)| = |\partial_v \mathrm{E}(u^*, v^*)| = 0.$$

**Remark 2.4.7.**

- *Finite dimensionality of* $\mathbb{U}$ *and* $\mathbb{V}$ *accounts for what is referred to as 'Assumption 3' by Ochs et al. (2019) and is some minimal condition which ensures that the limit is also a stationary point of our iteration (Equations* (2.10) *and* (2.11)*).*
- *The condition for finite dimensionality, as we shall see, does not directly relate to the non-convexity. The difficulty of showing convergence to critical points in infinite dimensions is common across both convex (Chambolle and Pock, 2011) and non-convex (Pock and Sabach, 2016; Ochs et al., 2019) optimisation.*

*Proof.* First we recall that, in finite dimensional spaces, convex functions are continuous on the relative interior of their domain (Duchi, 2017). Also note that by our choice of $\tau_u$ in Lemma 2.4.6, for all $u, u', v$ we have

$$\begin{aligned}|\,\mathrm{g}(J(u,v) + \nabla_u J(u,v)(u'-u)) - \mathrm{g}(J(u',v))| &\\ \leq |\,\mathrm{g}([J(u,v) - J(u',v)] - \nabla_u J(u,v)(u'-u))| &\\ = |\,\mathrm{g}(\nabla_u J(\widetilde{u},v)(u'-u) - \nabla_u J(u,v)(u'-u))| &\\ \leq \tau_u \left\|\widetilde{u}-u\right\|_{\mathbb{U}} \left\|u'-u\right\|_{\mathbb{U}} &\\ \leq \tau_u \left\|u'-u\right\|_{\mathbb{U}}^2 &\end{aligned}$$

where existence of such $\widetilde{u}$ is given by the mean value theorem. Hence, for all $u \in \mathbb{U}$ we have

$$\begin{aligned}\mathrm{E}(u_{n+1}, v_n) &= \mathrm{f}_{n+1,n} + \mathrm{g}(J_{n+1,n})\\ &\leq \mathrm{f}_{n+1,n} + \mathrm{g}(J_{n,n} + \nabla_u J_{n,n}(u_{n+1} - u_n)) + \tau_u \left\|u_{n+1} - u_n\right\|_{\mathbb{U}}^2\\ &\leq \mathrm{f}(u, v_n) + \mathrm{g}(J_{n,n} + \nabla_u J_{n,n}(u - u_n)) + \tau_u \left\|u - u_n\right\|_{\mathbb{U}}^2\\ &\leq \mathrm{f}(u, v_n) + \mathrm{g}(J(u, v_n)) + 2\tau_u \left\|u - u_n\right\|_{\mathbb{U}}^2\\ &= \mathrm{E}(u, v_n) + 2\tau_u \left\|u - u_n\right\|_{\mathbb{U}}^2\end{aligned}$$

where the first and third inequality are due to the condition shown above and the second is due to the definition of $u_{n+1}$ in (2.10). Finally, by continuity of f, $J$ and g we can take limits on both sides of this inequality:

$$\mathrm{E}(u^*, v^*) \leq \mathrm{E}(u, v^*) + 2\tau_u \left\|u - u^*\right\|_{\mathbb{U}}^2 \qquad \text{for all } u \in \mathbb{U}. \tag{2.16}$$

This completes the proof for $u^*$ as

$$|\partial_u \mathrm{E}(u^*, v^*)| = \limsup_{u \to u^*} \max\left(0, \frac{\mathrm{E}(u^*, v^*) - \mathrm{E}(u, v^*)}{\left\|u^* - u\right\|_{\mathbb{U}}}\right) \leq \limsup_{u \to u^*} 2\tau_u \left\|u - u^*\right\|_{\mathbb{U}} = 0.$$

The proof for $v^*$ follows by symmetry. □

**Remark 2.4.8.**

- *The important line in this proof, and where we need finite dimensionality, is being able to pass to the limit for* (2.16)*. In the general case we can only expect to have* $(u_n, v_n) \rightharpoonup (u^*, v^*)$*, typically guaranteed by choice of regularisation functionals as in our Theorem 2.3.1. In this reduced setting the left hand limit of* (2.16) *still remains valid,*

$$\mathrm{E}(u^*, v^*) \leq \liminf_{n\to\infty} \mathrm{E}(u_{n+1}, v_n) \text{ by weak lower semi-continuity.}$$

  *However, on the right hand side we require:*

$$\lim_{n\to\infty} \mathrm{E}(u, v_n) + 2\tau_u \|u - u_n\|_{\mathbb{U}}^2 \leq \mathrm{E}(u, v^*) + 2\tau_u \|u - u^*\|_{\mathbb{U}}^2 .$$

  *In particular, we already require* $\|u - u_n\|_{\mathbb{U}}$ *to be weakly upper semi-continuous. Topologically, this is the statement that weak and norm convergence are equivalent which will fail in most practical (infinite dimensional) examples.*

- *The properties we derive for* $(u^*, v^*)$ *are actually slightly stronger than that of Definition 2.4.4 which only depends on an infinitesimal ball about* $(u^*, v^*)$*. However,* (2.16) *gives us a quantification for the more global optimality of this point. This is pictured in Figure 2.7.*

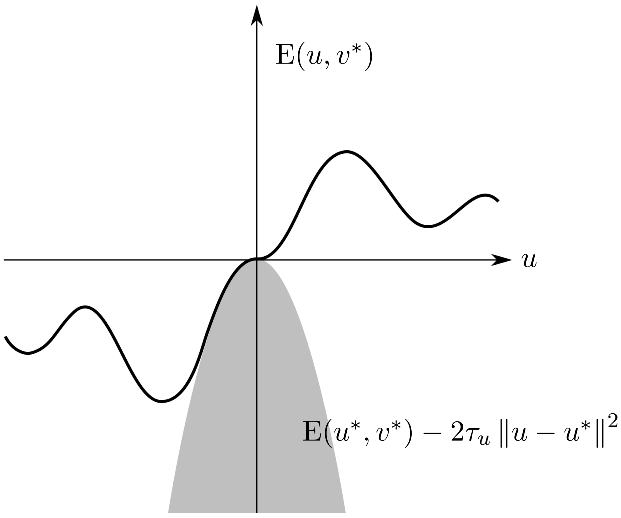


**Figure 2.7** Theorem 2.3.1 tells us that $(u^*, v^*)$ is a local critical point but does not qualify the globality of the limit point. Equation (2.16) further allows us to quantify the idea that if a lower energy critical point exists, then it must lie far from $(u^*, v^*)$. In particular, it must lie outside of the shaded cone given by the supporting quadratic.

### 2.4.4 Proof of Theorem 2.4.1

*Proof.* Fix arbitrary $(u_0, v_0) \in \mathbb{U} \times \mathbb{V}$ and $\tau_u, \tau_v \geq 0$. Let $(u_n, v_n)$ be defined as in Algorithm 2.1. By Lemma 2.4.3 we know that $\{(u_n, v_n) \text{ s. t. } n \in \mathbb{N}\}$ is contained in a sublevel set of E which in turn must be weakly compact by (2.5). The assumption of Lemma 2.4.6 is that we are in a finite dimensional space, therefore weak compactness is equivalent to strong compactness, i.e. some subsequence of $(u_n, v_n)$ converges in norm. Also by Lemma 2.4.6 we know that the limit point of this sequence must be an appropriate critical point. □

## 2.5 Results

For numerical results we present two synthetic examples and one experimental dataset from electron tomography. The two synthetic examples are discretised at a resolution of 200 × 200, then simulated using the X-ray transform with a parallel beam geometry sampled at 1° intervals over a range of 60° resulting in a full sinogram of size 287 × 180 and sub-sampled data at 287 × 60. Gaussian white noise (standard deviation 5% maximum signal) is then added to give the synthetic datasets. The experimental dataset was acquired with an annular dark field (parallel beam) Scanning TEM modality from which we have 46 projections spaced uniformly in 3° intervals over a range of 135°. Because of the geometry of the acquisition, we can treat the original 3D dataset as a stack of 2D sinograms and thus extract one of these slices as our example. This 2D dataset is then sub-sampled to 29 projections over 87°, reducing the size from 173 × 45 to 173 × 29. This results in a reconstruction with $u$ of size 120 × 120 and $v$ of size 173 × 180. A more detailed description of the acquisition and sample properties of the experimental dataset can be found in (Collins et al., 2015). The code, and data, for all examples is available[1] under the Creative Commons Attribution (CC BY) license.

### 2.5.1 Numerical details

All numerics are implemented in MATLAB 2016b. The sub-problem for $u$ is solved with a PDHG algorithm (Chambolle and Pock, 2011), while the sub-problem for $v$ is solved using the MOSEK solver via CVX (Mosek ApS, 2010; Grant and Boyd, 2014, 2008), the step size $\tau_u$ is adaptively calculated. The initial point of our algorithm is always chosen to be a good TV reconstruction, i.e.

$$u_0 = \operatorname*{argmin}_{u \geq 0} \frac{1}{2} \|\mathcal{S}\mathcal{R}u - \eta\|_2^2 + \mu \operatorname{TV}(u), \qquad v_0 = \mathcal{R}u_0.$$

[1] https://github.com/robtovey/2018_Directional_Inpainting_for_Limited_Angle

For clarity, we shall restate our full model with all of the parameters it includes. We seek to minimise the functional (2.4):

$$\mathrm{E}(u,v) = \frac{1}{2}\left\|\mathcal{R}u - v\right\|_{\alpha_1}^2 + \frac{\alpha_2}{2}\left\|\mathcal{S}\mathcal{R}u - \eta\right\|_2^2 + \frac{\alpha_3}{2}\left\|\mathcal{S}v - \eta\right\|_2^2 + \beta_1\,\mathrm{TV}(u) + \beta_2\left\|\mathcal{B}_{\mathcal{R}u}\nabla v\right\|_{2,1}$$

$$\mathcal{B}_\nu(\boldsymbol{x}) = c_1(\boldsymbol{x}|\lambda_1-\lambda_2,\lambda_1+\lambda_2)\boldsymbol{e}_1(\boldsymbol{x})\boldsymbol{e}_1(\boldsymbol{x})^\top + c_2(\boldsymbol{x}|\lambda_1-\lambda_2,\lambda_1+\lambda_2)\boldsymbol{e}_2(\boldsymbol{x})\boldsymbol{e}_2(\boldsymbol{x})^\top$$
$$\text{where } (\nabla\nu_\rho\nabla\nu_\rho^\top)_\sigma = \lambda_1\boldsymbol{e}_1\boldsymbol{e}_1^\top + \lambda_2\boldsymbol{e}_2\boldsymbol{e}_2^\top \text{ is a pointwise eigenvalue decomposition}$$
$$c_1(\boldsymbol{x}|\Delta,\Sigma) = 10^{-6} + \frac{\tanh(\Sigma(\boldsymbol{x}))}{1+\beta_3\Delta(\boldsymbol{x})^2},\quad c_2(\boldsymbol{x}|\Delta,\Sigma) = 10^{-6} + \tanh(\Sigma(\boldsymbol{x})).$$

We chose these particular $c_i$ according to two simple heuristics. If $\Sigma$ is large (steep gradients), then it is likely a region with edges and so the regularisation should be largest but still bounded above. If $\Delta = 0^+$, then there is a small or blurred 'edge' present and so we want to encourage it to become a sharp jump, i.e. $\nabla_\Delta c_1 < 0$. Theorem 2.2.1 tells us that choosing $c_i$ as functions of $\Delta^2$ will guarantee accordance with our later convergence results; this leads to our natural choice above. The number of iterations for Algorithm 2.1 was chosen to be 200 and 100 for the synthetic and experimental datasets respectively. To simplify the process of choosing values for the remaining hyper-parameters we made several observations:

1. The choice of $\alpha_i$ and $\beta_i$ appeared to be quite insensitive about the optimum. It is clear within 2-3 iterations whether values are of the correct order of magnitude. After this, values were only tuned coarsely. For example, $\alpha_3$ and $\beta_i$ are optimal within a factor of $10^{\pm 1/2}$.

2. We can chose $\alpha_2 = 1$ without any loss of generality. In which case, in general, $\beta_1$ should the same order of magnitude as when performing the TV reconstruction to get $u_0, v_0$.

3. $\alpha_2$ pairs $u$ to the given data and $\alpha_1$ pairs $u$ to the inpainted data, $v$. As such, $\alpha_1$ is spatially varying but should be something like a distance to the non-inpainting region. We chose the binary metric so that $u$ is paired to $v$ uniformly on the inpainting region and not at all outside.

4. DTV specific parameters $(\beta_2, \beta_3, \rho, \sigma)$ can be chosen outside of the main reconstruction. These were chosen by solving the toy problem:

$$\operatorname{argmin} \frac{1}{2}\left\|v - v_0\right\|_2^2 + \beta_2\left\|A_{v_0}\nabla v\right\|_{2,1}$$

   which is a lot faster to solve. $\rho > 0$ is required for the analysis and so this was fixed at 1. $\sigma$ is a length-scale parameter which indicates the separation between distinct edges. $\beta_3$ relies on the normalisation of the data. As can be seen in Table 2.1, for the two synthetic examples, with same discretisation and scaling, these values are also consistent. The only

| | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\rho$ | $\sigma$ |
|---|---|---|---|---|---|---|---|---|
| Concentric Rings Phantom | $\frac{1}{2^2}\mathbb{1}_{\Omega'^c}$ | 1 | $1\times10^{-1}$ | $3\times10^{-5}$ | $3\times10^{3}$ | $10^{10}$ | 1 | 8 |
| Shepp-Logan Phantom | $\frac{1}{4^2}\mathbb{1}_{\Omega'^c}$ | 1 | $3\times10^{-1}$ | $3\times10^{-5}$ | $3\times10^{2}$ | $10^{10}$ | 1 | 8 |
| Experimental Dataset (both sampling ratios) | $\frac{1}{2^2}\mathbb{1}_{\Omega'^c}$ | 1 | $3\times10^{2}$ | $1\times10^{-5}$ | $3\times10^{1}$ | $10^{6}$ | 1 | 0 |

**Table 2.1** Parameter choices for numerical experiments. Each algorithm was run for 300 iterations

value which changes is $\beta_2$, as expected, which weights how valid the DTV prior is for each dataset.

It is unclear whether a gridsearch may provide better results although, due to the number of parameters involved, this would definitely take a lot longer and mask some interpretability of the parameters. A further comparison of different choices of the main parameters can be found in Chapter A.

### 2.5.2 Canonical synthetic dataset

The first example shows two concentric rings. This is the canonical example for our model because the exact sinogram is perfectly radially symmetric which should trivialise the directional inpainting procedure, even with noise present. As can clearly be seen in Figure 2.8, the TV reconstruction is poor in the missing wedge direction which can be seen as a blurring out of the sinogram. By enforcing better structure in the sinogram, our proposed joint model is capable of extrapolating these local structures from the given data domain to recover the global structure and gives an accurate reconstruction.

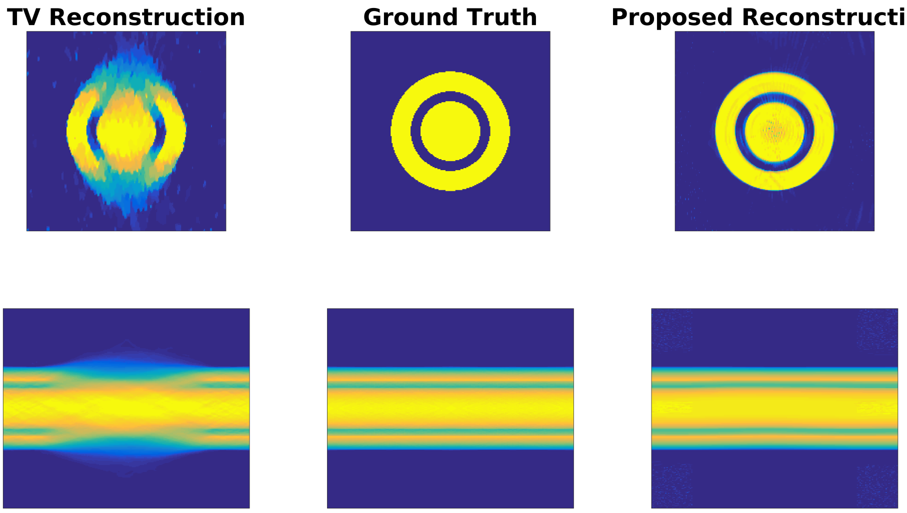


**Figure 2.8** Canonical synthetic example. Top row shows the reconstructions, $u$, while the bottom row shows the reconstructed sinogram, $v$.

### 2.5.3 Non-Trivial synthetic dataset

This example shows the modified Shepp-Logan phantom which is built up as a sum of ellipses. This example has a much more complex than before, although the sinogram still has a clear geometry. In Figure 2.9 we see that the largest scale feature, the shape of the largest ellipse, is recovered in our proposed reconstruction with minimal loss of contrast in the interior. One artifact we have not been able to remove is the two rays extending from the top of the reconstructed sample. Looking more closely we found that it was due to a small misalignment of the edge at the bottom of the sinogram as it crosses between the data to the inpainting region. Numerically, this happens because of the convolutions which take place inside the directional TV regularisation functional. Having a non-zero blurring is essential for regularity of the regularisation (Theorem 2.2.1) but the effect of this is that it does not heavily penalise misalignment on such a small scale. This means that at the interface between the fixed data-term there is a slight kink, the line is continuous but not $C^1$. The effect of this on the reconstruction is the two lines which extend from the sample at this point. Looking at quantitative measures, the PSNR value rises from 17.33 to 17.36 whereas the SSIM decreases from 0.76 to 0.62, from TV to the proposed reconstruction, respectively. These measures are inconclusive and the authors feel that they fail to balance the improvement to global geometry verses more local artifacts in the reconstructions.

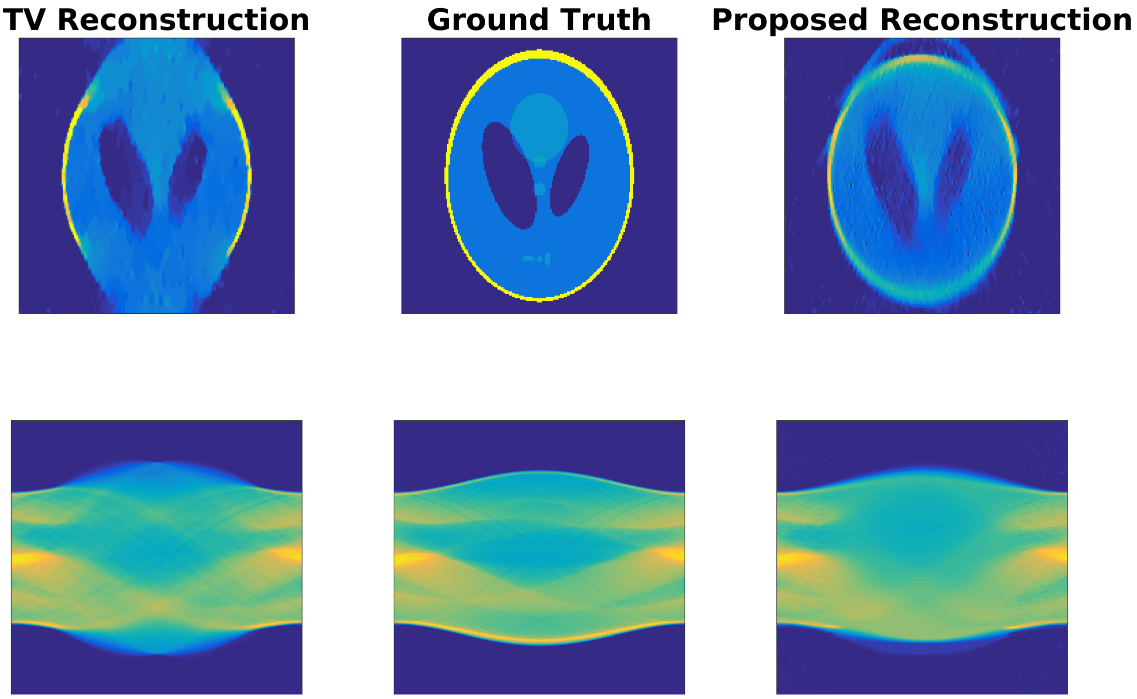


**Figure 2.9** Non-trivial synthetic example of the modified Shepp-Logan phantom. Top row shows the reconstructions, $u$, while the bottom row shows the reconstructed sinogram, $v$. We regain the large-scale geometry of the shape without losing much of the interior features.

### 2.5.4 Experimental dataset

The experimental sample is a silver bipyramidal crystal placed on a planar surface, and the challenges of the recorded dataset are shown in Figure 2.10. We immediately see that the wide angle projections have large artifacts which produces a very low signal to noise ratio. Another issue present is that there is mass seen in some of the projections which cannot be represented within the reconstruction volume. Both of these issues violate the simple X-ray model that is used. Exact modelling would require estimation of parameters which are not available a priori and so the preferred acquisition is one which automatically minimises these modelling errors. Another artifact is that over time each surface becomes coated with carbon. This is a necessary consequence of the sample preparation and this build-up is known to occur during the microscopy. The result of modelling errors and time dependent noise is to prefer an acquisition with limited angular range and earliest acquired projections. Because of this, in numerical experiments we compare both TV and our proposed reconstruction using only $^3\!/\!_4$ of the available data, 29 projections over an 87° interval, with a bias towards earlier projections. The artifacts due to the out-of-view mass are unavoidable, but we can perform some further pre-processing to minimise the effect. In particular, if we shrink the field of view of the detector, then the 'heaviest' part of the data will be the particle of interest and the model violations will be relatively small, increasing the signal to noise ratio. This can be seen as the sharp horizontal cut-off in the pre-processed sinograms seen on the right of Figure 2.10. The effect of this on the reconstruction is going to be that there is a thin ring of mass placed at the edge of the (shrunken) detector view which will be clearly identifiable in the reconstruction. As a ground truth approximation we shall use a TV reconstruction on the full data for the location of the boundaries of the particle, alongside prior knowledge of this sample for more precise geometrical features. We also note that the particle should be very homogeneous so this is another example where we expect the TV reconstruction to be very good.

The sample is a single crystal of silver and so we know it must have very uniform intensity and we are interested in locating the sharp facets which bound the crystal (Collins et al., 2015). In Figure 2.11 we immediately see that the combination of homogeneity and sharp edges is better reconstructed in our proposed reconstruction. Because we expect the reconstruction to be constant on the background and the particle, thresholding the reconstruction allows us to easily locate the boundaries and estimate interior angles of the particle. Figure 2.12 shows such images where the threshold is chosen to be the approximate midpoint of these two levels. We see that the proposed reconstruction consistently has less jagged edges and the left-hand corner is better curved, as is consistent with our knowledge of the sample. Looking back at the full colour images we see that this is a result of lack of sharp decay at the boundary and homogeneity inside the sample. Looking for boundary location error, we see the biggest error in both TV and joint reconstruction is on the bottom-left edge where both reconstructions pull the line inwards. However, looking particularly at points $(40, 80)$ and $(20, 60)$, we see that this was less severe in the proposed method. The other missing wedge artifact is in the top-right corner which has been extended in both reconstructions although it is thinner in the proposed reconstruction. This indicates that it was better able to continue the straight edges either side of the corner and the blurring in the missing wedge direction is more localised than in the TV reconstruction. Overall, we see see that the proposed reconstruction method is much more robust to a decrease in angular sampling range.

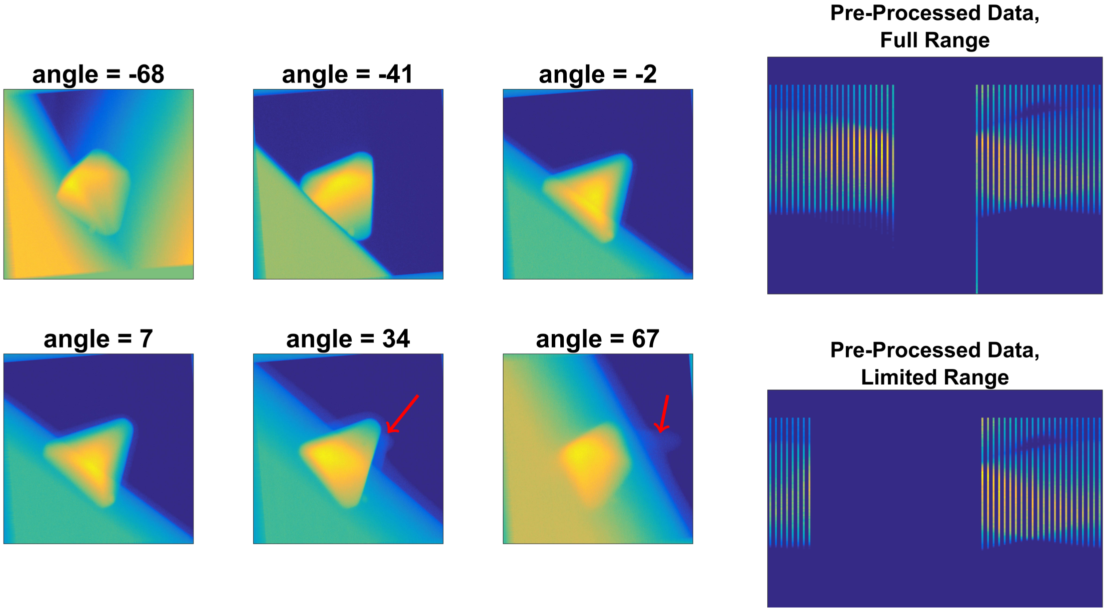


**Figure 2.10** Raw data for EM example. Projections at large angles, e.g. −68°, show the presence of the sample holder which violates the X-ray modelling assumption that outside of the region of interest is vacuum. If the violation is too extreme, then this can cause strong artifacts in reconstructions and so the common action is to discard such data. The plane surface also violates this model but is relatively weak at low angles and so will cause weaker artifacts. A source of noise in this acquisition is that over time the surface becomes coated with carbon. This is first visible as a thin film at −2° and progressively gets thicker through the remaining projections. At 34° we see a bump of carbon appear on the top right edge. After pre-processing, we extract a 2D slice of all projections to form the full range as shown top right artificially sub-sample to compare TV with our proposed reconstruction method.

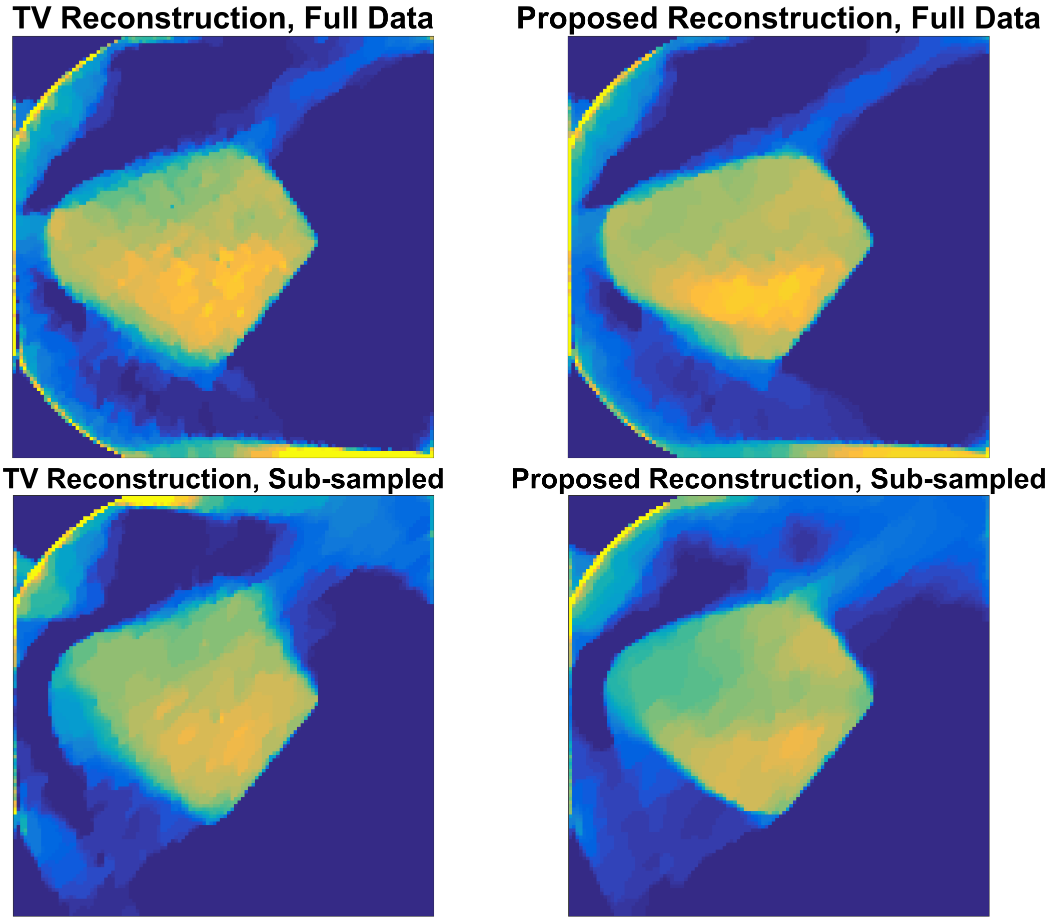

TV Reconstruction, Full Data
Proposed Reconstruction, Full Data
TV Reconstruction, Sub-sampled
Proposed Reconstruction, Sub-sampled

**Figure 2.11** Reconstructions from a slice of the experimental data. We have chosen the slice half-way down through the projections shown in Figure 2.10 to coincide with one of the rounded corners. The arc artifact was an anticipated consequence due to the out-of-view mass, the pre-processing has simply reduced the intensity. Proposed reconstructions consistently show better homogeneity inside the particle and sharper boundaries. The missing angles direction is the bottom-left to top-right diagonal where we see most error in each reconstruction, in particular, the blurring of the top right corner of the particle is a limited angle artifact.

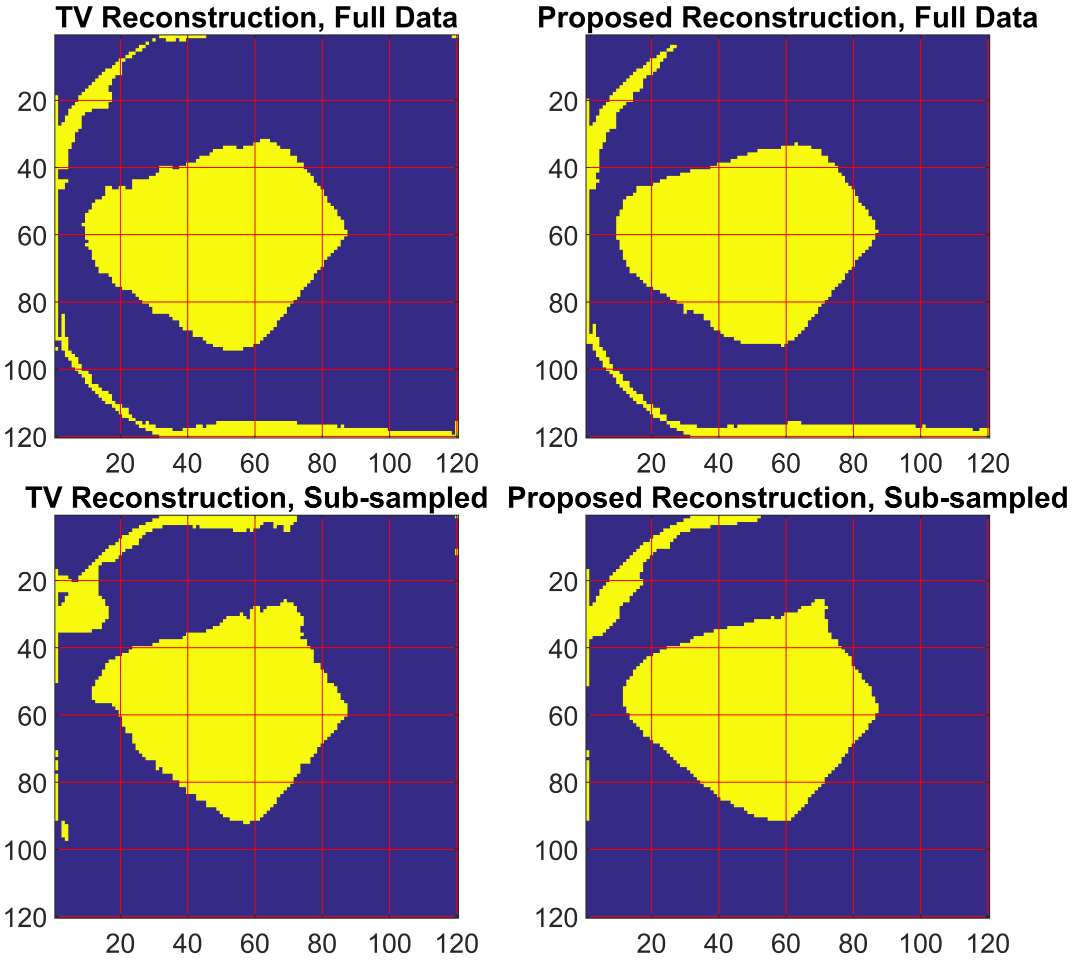


**Figure 2.12** Comparison between each reconstruction after thresholding. The geometrical properties of interest are that each edge should be linear, the left hand corner is rounded and the remaining corners are not. The particle of interest is homogeneous so thresholding the images should emphasise this in a way which is very unsympathetic to blurred edges. Again, the top right corner of each particle in the sub-sampled reconstructions coincides with the exacerbated missing wedge direction and so we expect each reconstruction to make some error here.

## 2.6 Conclusions and outlook

In this chapter we have presented a novel method for tomographic reconstructions in a limited angle scenario along with a new numerical algorithm with convergence guarantees. We have also tested our approach on synthetic and experimental data and shown consistent improvement over alternative reconstruction methods. Even when the X-ray transform model is noticeably violated, as with our experimental data, we still better recover boundaries of the reconstructed sample.

There are three main directions which could be explored in future. Firstly, we think there is great potential to apply our framework to other applications, such as in tomographic imaging with occlusions and heavy metal artifacts where the inpainting region is much smaller (Köstler et al., 2006; Zhang et al., 2011). Secondly, we would like to find an alternative numerical algorithm with either faster practical convergence, or one which is more capable of avoiding local minima in this non-convex landscape. Finally, we would like to explore the potential for an alternative regularisation functional on the sinogram which is better able to treat visible and invisible singularities, denoising and inpainting problems, independently. At the moment, the TV prior alone can reconstruct visible singularities well however, introducing a sinogram regulariser currently improves on the invisible region at the expensive of damaging the visible. Overall, we feel that this presents the natural progression for the current work, although it remains unclear how to regularise these invisible singularities.

# Chapter 3

# Strain Tomography of Crystals

Strain is the material property that corresponds to the measure of deformation of an object, described by a $3 \times 3$ tensor at every point in the volume. Strain engineering, creating objects with a chosen strain distribution, is common in a range of modern industries, but relies on the accurate measurement of strain to evaluate a manufacturing process. Direct nanoscale measurement of this tensor field inside these materials has been limited by both a lack of experimental and analytical tools. Scanning electron diffraction has emerged as a powerful tool for reconstructing two-dimensional maps of ‘average strain’ for samples with simple structure. The obstacle to generalising this technique to full three-dimensions has been a lack of a formal framework for understanding the averaging process for general strain fields.

In this chapter, we propose a framework and analyse the inverse problem for three-dimensional reconstruction of the full strain tensor field. There are two analytical complexities arising from non-linearities in the first Born approximation:

1. diffraction intensity is proportional to the square of the wave function,

2. the wave function is not linear with respect to strain parameters.

These are the main issues to be overcome when developing a linear tomography problem for strain mapping. Our proposed linear model is analytically motivated and shown to be numerically accurate. This shows that strain can be recovered as an ill-posed linear tensor tomography inverse problem with missing and corrupt data. Numerical results show that this inverse problem can also be solved accurately with realistic data by utilising total variational regularisation.

## 3.1 Related work

Nanoscale strain is widely used to engineer desirable materials properties, for example, improving field effect transistor (FET) performance (Chu et al., 2009), opening a bulk bandgap in topological insulator systems (Hsieh et al., 2009) and enhancing ferroelectric properties (Schlom et al., 2007). Strain also arises around crystal defects, which further affect materials properties. 3D strain reconstruction of one or more strain components has been achieved using X-ray diffraction techniques, including: coherent Bragg diffractive imaging (Pfeifer et al., 2006; Robinson and Harder, 2009; Newton et al., 2010), micro-Laue diffraction using a differential aperture (Larson et al., 2002), and diffraction from polycrystalline specimens combined with back-projection methods (Korsunsky et al., 2006, 2011). The spatial resolution of these X-ray techniques is however limited to ca. 20-100 nm and sub-10 nm resolution strain mapping is therefore dominated by (scanning) TEM techniques (Hÿtch and Minor, 2014). These techniques include: imaging at atomic resolution (Hÿtch et al., 2003; Galindo et al., 2007), electron holography (Hÿtch et al., 2011) and SED (Usuda et al., 2005; Béché et al., 2013). Amongst these, 3D strain has been assessed by atomic resolution tomography (Goris et al., 2015) and in a proof-of-principle reconstruction of a single strain component using SED (Johnstone et al., 2017). In 2D strain mapping, SED has emerged as a particularly versatile and precise approach to strain mapping with few nanometre resolution (Cooper et al., 2015, 2017).

As was shown in Figure 1.6, SED is a 4D-STEM technique (Ophus, 2019) based on the acquisition of a 2D transmission electron diffraction pattern at every probe position as a focused electron probe is scanned across the specimen in a 2D scan. It was also shown in Section 1.4.4, particularly Figure 1.9, that diffraction patterns from crystalline samples are comprised of sparse spikes (formally Bragg discs) of intensity at predictable locations. The principle of strain mapping of crystalline materials using SED is that the movement of spikes can be interpreted as components of the 2D strain (Usuda et al., 2005). Over a 2D scan, this becomes a spatial map of (2D) strain.

Strain maps of three path averaged components of the strain tensor in 2D have been reported from a wide range of materials via SED (Cooper et al., 2017; Haas et al., 2017; Pekin et al., 2018; Bonef et al., 2016). Determining the position of the Bragg disks in each diffraction pattern is a critical step, which has been explored in recent literature with cross-correlation based disk finding approaches achieving the best accuracy and precision (Béché et al., 2009; Rouviere et al., 2013; Pekin et al., 2017; Zeltmann et al., 2020). The incorporation of double-conical electron beam rocking (see Section 1.4.5), to record scanning precession electron diffraction (SPED) data, has further been demonstrated to improve precision both numerically (Mahr et al., 2015) and experimentally (Cooper et al., 2015). However, progress towards 3D strain mapping using S(P)ED data has so far been limited to a proof-of-principle reconstruction of one strain component in 3D (Johnstone et al., 2017) by the lack of a framework for three-dimensional strain tensor field reconstruction using S(P)ED data. In this work, we establish such a framework for three-dimensional strain tensor field reconstruction from S(P)ED data via consideration of an analytical forward model (see Sections 3.3 and 3.4). We show that a linearised approximation coincides with a non-symmetric tensor tomography problem and use this to demonstrate recovery of the full strain field.

In Section 3.6 we show that the linearised model for diffraction data coincides with the transverse ray transform (TRT) introduced in Section 1.4.8. Finally, we validate our reasoning computationally in Section 3.7 with a range of complex forward models for scanning electron diffraction. Our analytical and numerical results establish a robust framework for three-dimensional strain tensor field reconstruction using S(P)ED data.

## 3.2 Notation

In the scope of this chapter we will adopt the following notation.

- $D\colon \mathbb{R}^2 \to \mathbb{R}$ and subscripted derivatives denote a two-dimensional diffraction pattern recorded on a flat detector
- $\boldsymbol{K} = (\boldsymbol{k}, k_z) = (k_x, k_y, k_z)$ gives the standard coordinates on $\mathbb{R}^3$, in Fourier space (reciprocal space), which we treat as $\mathbb{R}^3 = \mathbb{R}^{2+1}$ interchangeably, where $\boldsymbol{k}$ is the 2D coordinate on the detector
- $\Gamma = \left(\begin{smallmatrix}1&0\\0&1\\0&0\end{smallmatrix}\right)$ is the natural lifting from $\mathbb{R}^2$ to the plane $k_z = 0$ in $\mathbb{R}^3$. i.e. $\Gamma\boldsymbol{K} = \boldsymbol{k}$, $\Gamma^\top\boldsymbol{k} = (\boldsymbol{k}, 0)$

## 3.3 Principles of strain diffraction imaging

The continuum definition of strain is a direct measure of deformation of an object. The technical definition is as follows.

**Definition 3.3.1.** *Let $u\colon \mathbb{R}^3 \to \mathbb{R}$ define an electrostatic potential and $\vec{R}\colon \mathbb{R}^3 \to \mathbb{R}^3$ define a deformation by*

$$u'(\boldsymbol{r}) = u(\boldsymbol{r} + \vec{R}(\boldsymbol{r})).$$

*where $\vec{R}$ is the* displacement map *and $\vec{\mathrm{E}} = \nabla\vec{R}$ is the* displacement gradient tensor field. Strain *is the symmetric component of the displacement gradient tensor. In this work we adopt the convention*

$$\vec{\varepsilon} = \tfrac{1}{2}(\vec{\mathrm{E}} + \vec{\mathrm{E}}^\top).$$

While the definition of the displacement gradient tensor is un-ambiguous, usage of the word *strain* is much looser. Colloquially they can be used synonymously however another common formal definition is

$$\vec{\varepsilon} = \sqrt{\vec{\mathrm{E}}^\top\vec{\mathrm{E}}}$$

i.e. the geometric mean instead of the arithmetic mean. We choose to go with the former definition as it aligns with that in the tensor tomography literature. Since diffraction patterns are sensitive to both symmetric and nonsymmetric deformations, we will provide analysis for the full deformation gradient tensor which is equally valid for both conventions of strain.

### 3.3.1 Recap of diffraction imaging

In this work we build on the main results of Sections 1.4.2 to 1.4.5. We restate the key results here for clarity.

The diffraction model used for analysis in this chapter will be the kinematical Ewald sphere model. In Equation (1.5) this is stated as

$$D(\boldsymbol{k}) = \left|\mathcal{F}[\Psi_p u]\right|^2(\boldsymbol{k}, k_z(\boldsymbol{k})), \qquad k_z(\boldsymbol{k}) := 2\pi\lambda^{-1} - \sqrt{4\pi^2\lambda^{-2} - |\boldsymbol{k}|^2} \tag{3.1}$$

where $\boldsymbol{k}$ is the 2D coordinate on the detector, $\Psi_p$ is the probe function and the incident electron beam is chosen to be travelling along the $z$-axis.

We will also rely on the technique of beam precession which results in the expression of Equation (1.8):

$$\begin{aligned} D_\alpha(\boldsymbol{k}) = \mathop{\mathbb{E}}_t \Bigg\{ &\left|\mathcal{F}[\Psi_p u(R_t x)]\right|^2(\boldsymbol{k}, k_z(\boldsymbol{k})) \text{ such that } t \in [0, 2\pi) \text{ and} \\ &R_t = \begin{pmatrix} \cos(t) & \sin(t) & 0 \\ -\sin(t) & \cos(t) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) \\ 0 & -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} \cos(t) & -\sin(t) & 0 \\ \sin(t) & \cos(t) & 0 \\ 0 & 0 & 1 \end{pmatrix} \Bigg\} \end{aligned} \tag{3.2}$$

where $\alpha$ is the precession angle.

We use a kinematical model for analytical results but also the multislice dynamical model for numerical results in Section 3.7. The key physical difference, as desccribed in Section 1.4.3, is that the dynamical model accounts for each electron to experience multiple scattering events. The kinematical model is still qualitatively accurate but provides a much simpler analytical form.

In a similar vein, in the crystallography literature, Definition 3.3.1 is referred to as the *deformable ion* model as it allows the shape of atoms to be deformed. In the discrete world of atomic lattices, the alternative is the *rigid ion* model where only the atom centre is 'deformed' or translated by $\vec{R}$ (Howie and Basinski, 1968). We will use the deformable ion model for analytical argument but the rigid ion model for numerical work as it is thought to be more accurate.

### 3.3.2 Technical assumptions for diffraction imaging

In Section 1.4.4 we saw both equations and pictures to describe diffraction for crystals. Both indicate that diffraction patterns are a sparse sum of sharp peaks but we are yet to make this link explicit. Even before considering strain, additional assumptions are require to justify the precise behaviour.

More concretely, examples of diffraction simulations were shown in Figure 1.9 (reproduced in Figure 3.1 for convenience) and Lemma 1.4.3 computed the analytical form for a truncated crystal. In particular, if $u_0$ is an ideal, infinite conventional crystal and $u$ is described by

$$u(x) := \mathbb{1}_{|\cdot|_\infty \leq \rho/2} u_0(x) = \begin{cases} u_0(x) & x \in [-\rho/2, \rho/2]^3 \\ 0 & \text{else} \end{cases}$$

then the (kinematical) diffraction pattern can be written as

$$D(\boldsymbol{k}) = \left| \sum_{i=1}^{\infty} a_i \rho \operatorname{sinc}(\rho[k_z(\boldsymbol{k}) - p_{i,z}]) \, \mathrm{f}(\boldsymbol{k} - \Gamma^\top \boldsymbol{p}_i) \right|^2$$

where

$$\mathrm{f}(\boldsymbol{k}) := \rho^2 \mathcal{F}[\Psi_p] \star [\operatorname{sinc}(\rho \cdot)](\boldsymbol{k}) = \rho^2 \int_{\mathbb{R}^2} \mathcal{F}[\Psi_p](\boldsymbol{k} - \boldsymbol{k}') \operatorname{sinc}(\rho \boldsymbol{k}') d\boldsymbol{k}'.$$

For the remainder of this work we make the following assumptions.

*thick crystals*, i.e. $\rho \sim 1000\,\text{Å}$ (Assumption 1)

*small wavelength*, i.e. $\lambda \sim 0.01\,\text{Å}$ (Assumption 2)

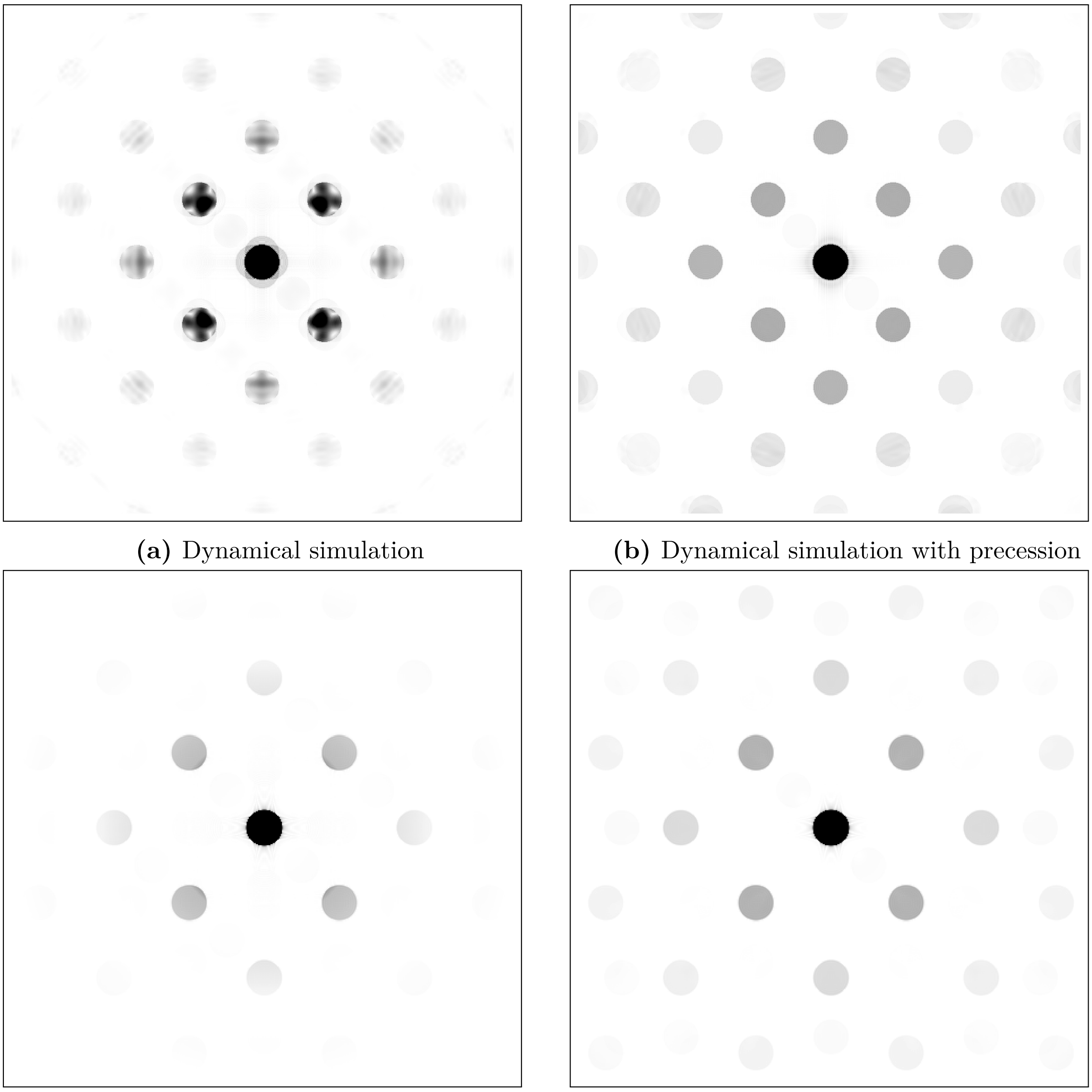
*non-overlapping Bragg disks*, i.e. $\exists \bar{r} > 0$ such that for all $i_1 \neq i_2$ s. t.

$$|\boldsymbol{p}_{i_1} - \boldsymbol{p}_{i_2}| > 2\bar{r} \text{ and } |\boldsymbol{k}| > \bar{r} \implies \mathrm{f}(\boldsymbol{k}) = 0 \qquad \text{(Assumption 3)}$$

*symmetrical spots*, i.e. $\mathrm{f}(-\boldsymbol{k}) = \mathrm{f}(\boldsymbol{k})$ for all $\boldsymbol{k}$ (Assumption 4)

*narrow beam*, i.e. $|\boldsymbol{r}| > c \sim 30\,\text{Å} \implies \Psi_p(\boldsymbol{r}) = 0$ (Assumption 5)

These assumptions can all be met readily in typical SPED experiments by configuring the electron optics while considering the crystal lattice parameters. Informally, Lemma 1.4.3 showed that diffraction patterns are a sum of spikes with shape f, at locations $\Gamma^\top \boldsymbol{p}_i = (p_{i,x}, p_{i,y})$ and with intensity dependent on $p_{i,z}$. Assumption 1 justifies that the only visible spikes are those with $p_{i,z} \approx k_z(\boldsymbol{k})$ and Assumption 2 refines this to $p_{i,z} \approx 0$. Because of the lattice structure of conventional crystals (see Definition 1.4.1), if the $z$-direction coincides with a *zone axis* of the crystal then we can say $p_{i,z} \approx 0 \iff p_{i,z} = 0$. Assumption 3 and Assumption 4 state that the spikes are symmetrical and well separated. For the kinematical model, Assumption 4 is equivalent to the symmetry of $\Psi_p$. Finally, Assumption 5 dictates the spatial resolution of the strain map reconstruction. The width of the support of $\Psi_p$ acts like a point-spread function on the model. Without treating this explicitly, we need to assume that the beam is contained within a single column of voxels in the final strain mapped volume.



**(a)** Dynamical simulation

**(b)** Dynamical simulation with precession

**(c)** Kinematical simulation

**(d)** Kinematical simulation with precession

**Figure 3.1** (Duplication of Figure 1.9.) Simulations of diffraction patterns from an unstrained Silicon crystal. (a) shows a dynamical simulation where complex spot inhomogeneities can be seen. (b) shows that with precession, the intensities in the multislice simulation become much more homogeneous. (c)/(d) show kinematical simulations without/with precession. Note that precessed images qualitatively agrees very closely with each other.

### 3.3.3 Strained Fourier transforms

The kinematical model highlights the principle that we can understand diffraction by understanding the Fourier transform. The following theorem considers the case of a crystal subject to uniform affine deformation.

**Theorem 3.3.2.** *If*

$$u'(\boldsymbol{r}) = u(\boldsymbol{r} + \vec{R}(\boldsymbol{r})) = u(A\boldsymbol{r} + \boldsymbol{b}) \textit{ for some } A \in \mathbb{R}^{3\times 3}, \boldsymbol{b} \in \mathbb{R}^3, u \in L^2(\mathbb{R}^n; \mathbb{C}),$$

*where $A$ is invertible, then we can express its Fourier transform as:*

$$\mathcal{F}[u'](\boldsymbol{K}) = \det(A)^{-1} e^{\imath \boldsymbol{b} \boldsymbol{\cdot} A^{-\top}\boldsymbol{K}} \mathcal{F}[u](A^{-\top}\boldsymbol{K}).$$

This is a standard result in Fourier analysis and a proof is given in Section B.1. The appearance of the determinant emphasises that larger potentials diffract more strongly. The only changes to the Fourier transform are a corresponding linear deformation and a change of phase depending on the translation. In Section 1.4.4 it was of interest to consider idealised crystals, say $u_0$ such that

$$\mathcal{F}[u_0] = \sum_{i=1}^{\infty} a_i \delta_{\boldsymbol{p}_i}.$$

In this case the Fourier transform is a distribution and so requires an additional technical lemma.

**Lemma 3.3.3.** *If $A \in \mathbb{R}^{3\times 3}$ is an invertible matrix and $\boldsymbol{b} \in \mathbb{R}^3$ then*

$$\mathcal{F}[u_0(A\cdot + \boldsymbol{b})](\boldsymbol{K}) = e^{\imath \boldsymbol{b} \boldsymbol{\cdot} A^{-\top}\boldsymbol{K}} \sum_{i=1}^{\infty} a_i \delta_{A^{\top}\boldsymbol{p}_i}(\boldsymbol{K}).$$

Due to the importance of this standard result in this work we include a proof in Section B.1. This lemma demonstrates a key feature for diffraction from deformed single crystals. Under linear deformation, the location of spikes in Fourier space is equivalently linearly deformed and therefore the locations of diffracted peaks are also linearly deformed.

### 3.3.4 Strained diffraction patterns

To model the diffraction patterns produced by crystals subject to more general deformation we consider a crystal subject to piecewise affine deformation. This is an implicit assumption that deformed crystals are also piecewise smooth and so can be well approximated in this framework. We thus assume the material, $u$, may be expressed as:

$$u(\boldsymbol{r}) = \sum_{j=1}^{N} u_0(A_j\boldsymbol{r} + \boldsymbol{b}_j) \mathbb{1}_{|\cdot|_\infty \leq 1/2} \left(\frac{\boldsymbol{r} - \boldsymbol{\beta}_j}{\rho}\right) \tag{3.3}$$

where:

$j \in [N] = \{1, \dots, N\}$ indexes over all, finitely many, voxels

$\boldsymbol{\beta}_j \in \mathbb{R}^3$ is the location of voxel $j$

$A_j \in \mathbb{R}^{3\times 3}$ is $A_j = \mathrm{id} + \overrightarrow{\mathrm{E}}$ within voxel $j$

$\boldsymbol{b}_j \in \mathbb{R}^3$ represents the shift to align $u_0$ with the $u$ in voxel $j$.

In this work, each voxel is a volume element that is sufficiently large that the discrete atoms blur into a continuum. We note that the results which follow can be extended to more generic tensor fields by limiting the voxel size to zero and replacing the finite sum with a Riemann integral if additional smoothness assumptions are made on $u_0$, $\boldsymbol{b}_j$, and $A_j$ to guarantee any necessary exchanges of limits.

We can now express the full diffraction pattern of a deformed crystal.

**Lemma 3.3.4.** *If the probe is narrow (Assumption 5) then*

$$\begin{aligned} D(\boldsymbol{k}) &= |\mathcal{F}[\Psi_p] \star \mathcal{F}[u]|^2(\boldsymbol{k}, k_z(\boldsymbol{k})) \\ &= \left| \sum_{i\in\mathbb{N}, \Gamma^\top \boldsymbol{\beta}_j = \boldsymbol{0}} \widehat{a}_i(k_z(\boldsymbol{k}) - (A_j^\top \boldsymbol{p}_i)_z, \beta_{j,z}) e^{\imath \boldsymbol{b}_j \bullet \boldsymbol{p}_i} \mathrm{f}(\boldsymbol{k} - \Gamma^\top A_j^\top \boldsymbol{p}_i) \right|^2 \end{aligned}$$

*where*

$$\widehat{a}_i(k, \beta) = a_i \rho \operatorname{sinc}(\rho k) e^{-\imath \beta k}.$$

*Proof.* By Theorem 3.3.2 we have

$$\mathcal{F}\left[\mathbb{1}_{|\cdot|_\infty \leq 1/2}\left(\tfrac{\boldsymbol{r}-\boldsymbol{\beta}_j}{\rho}\right)\right](\boldsymbol{K}) = \rho^3 \operatorname{sinc}(\rho \boldsymbol{K}) e^{-\imath \boldsymbol{\beta} \bullet \boldsymbol{K}} = \left[\rho^2 \operatorname{sinc}(\rho \boldsymbol{k}) e^{-\imath (\Gamma^\top \boldsymbol{\beta}) \bullet \boldsymbol{k}}\right] \left[\rho \operatorname{sinc}(\rho k_z) e^{-\imath \beta_z k_z}\right]$$

which can be split into its $(x, y)$ and $z$ components. Combining this again with Theorem 3.3.2 we can expand

$$\mathcal{F}[\Psi_p] \star \mathcal{F}[u] = \sum_{i \in \mathbb{N}, j \in [N]} \left( \mathcal{F}[\Psi_p] \star \mathcal{F}\left[\mathbb{1}_{|\cdot|_\infty \leq 1/2}\left(\frac{\boldsymbol{r} - \boldsymbol{\beta}_j}{\rho}\right)\right]\right) \star \left[a_i e^{\imath \boldsymbol{b}_j \bullet A_j^{-\top} \boldsymbol{K}} \delta_{A_j^\top \boldsymbol{p}_i}(\boldsymbol{K})\right].$$

To simplify this, observe that for all smooth functions, $\varphi$, $\psi$:

$$\begin{aligned} \varphi \star (\psi \delta_{\boldsymbol{p}})(\boldsymbol{K}) &= \int_{\mathbb{R}^3} \varphi(\boldsymbol{K} - \boldsymbol{K}') \psi(\boldsymbol{K}') \delta_{\boldsymbol{p}}(\boldsymbol{K}') d\boldsymbol{K}' \\ &= \varphi(\boldsymbol{K} - \boldsymbol{p}) \psi(\boldsymbol{p}). \end{aligned}$$

Thus we derive

$$\mathcal{F}[\Psi_p] \star \mathcal{F}[u](\boldsymbol{K}) = \sum_{i\in\mathbb{N}, j\in[N]} \left(\mathcal{F}[\Psi_p] \star \mathcal{F}\left[\mathbb{1}_{|\cdot|_\infty \leq 1/2}\left(\frac{\boldsymbol{r}-\boldsymbol{\beta}_j}{\rho}\right)\right]\right)(\boldsymbol{K} - A_j^\top \boldsymbol{p}_i)\left[a_i e^{\imath \boldsymbol{b}_j \bullet \boldsymbol{p}_i}\right]$$

Finally, if the beam is smaller than the width of a single block (from Assumption 5, $2r < \rho$) then only one column of blocks directly on the beam path contribute to the diffraction signal. Without loss of generality, this is the set of blocks $j$ such that $\beta_{j,x} = \beta_{j,y} = 0$, equivalently $\Gamma^\top \boldsymbol{\beta}_j = \mathbf{0}$. With this simplification, we can expand

$$\begin{aligned}
\mathcal{F}[\Psi_p] \star \mathcal{F}\left[\mathbb{1}_{|\cdot|_\infty \leq 1/2}\left(\frac{\boldsymbol{r}-\boldsymbol{\beta}_j}{\rho}\right)\right](\boldsymbol{K}) &= \left[\mathcal{F}[\Psi_p] \star [\rho^2 \operatorname{sinc}(\rho\boldsymbol{\kappa}) e^{-\imath(\Gamma^\top \boldsymbol{\beta}_j)\bullet\boldsymbol{\kappa}}]\right](\boldsymbol{k})\left[\rho \operatorname{sinc}(\rho k_z) e^{-\imath \beta_{j,z} k_z}\right] \\
&= \left[\mathcal{F}[\Psi_p] \star [\rho^2 \operatorname{sinc}(\rho\boldsymbol{\kappa})]\right](\boldsymbol{k})\left[\rho \operatorname{sinc}(\rho k_z) e^{-\imath \beta_{j,z} k_z}\right] \\
&\equiv \mathrm{f}(\boldsymbol{k})\left[\rho \operatorname{sinc}(\rho k_z) e^{-\imath \beta_{j,z} k_z}\right].
\end{aligned}$$

Substituting this above gives

$$\mathcal{F}[\Psi_p] \star \mathcal{F}[u](\boldsymbol{K}) = \sum_{i\in\mathbb{N}, \Gamma^\top\boldsymbol{\beta}_j = \mathbf{0}} a_i[\rho \operatorname{sinc}(\rho\cdot) e^{-\imath\beta_{j,z}\cdot}](k_z - (A_j^\top \boldsymbol{p}_i)_z)\; e^{\imath \boldsymbol{b}_j \bullet \boldsymbol{p}_i}\, \mathrm{f}(\boldsymbol{k} - \Gamma^\top A_j^\top \boldsymbol{p}_i)$$

as required. □

A key point here is that Assumption 5 is used to guarantee that the beam is completely contained within a single column of voxels. Without loss of generality, this column is centred at $x = y = 0$ which reduces the sum to indices $j$ such that $\Gamma^\top \boldsymbol{\beta}_j = \mathbf{0}$.

In the limit $\rho \to \infty, \lambda \to 0$ (large voxels, high energy incident electrons) Lemma 3.3.4 simplifies to

$$\lim_{\substack{\lambda\to 0,\\ \rho\to\infty}} \frac{D(\boldsymbol{k})}{\rho} = \left|\sum_{i,j\in I} a_i e^{\imath \boldsymbol{b}_j \bullet \boldsymbol{p}_i} \mathcal{F}[\Psi_p](\boldsymbol{k} - \Gamma^\top A_j^\top \boldsymbol{p}_i)\right|^2 \tag{3.4}$$

where $I = \{(i,j) \in \mathbb{N}^2 \text{ s.t. } \Gamma^\top \boldsymbol{\beta}_j = \mathbf{0},\ (A_j^\top \boldsymbol{p}_i)_z = 0\}$. This helps to highlight two key properties of diffraction imaging under deformation:

- 'in-plane' deformation moves the centre of the spot linearly from $(p_{i,x}, p_{i,y})$ to $\Gamma^\top A_j^\top \boldsymbol{p}_i = ((A_j^\top \boldsymbol{p}_i)_x, (A_j^\top \boldsymbol{p}_i)_y)$.
- 'out-of-plane' deformation changes the intensity of each spot depending on $(A_j^\top \boldsymbol{p}_i)_z$. In the high-energy limit, this is absolute however, in practice the intensity of spot $i$ is dampened by a factor of $\operatorname{sinc}(\rho(k_z(\boldsymbol{k}) - A_j^\top \boldsymbol{p}_i)_z) \leq 1$. In either case, dependence of the intensities on the out-of-plane strain $A_j$ is highly non-linear.

Lemma 3.3.4 provides a very explicit model for computing diffraction patterns from deformed crystals yet it is still too complex to directly derive a linear correspondence for the strain mapping inverse problem. To do this we will use precession and also one final technical assumption, that we are in a ‘small strain’ scenario. Formally, we state this as:

$$\text{small strain, i.e. } |A_j - \mathrm{id}| < \sigma \ll 1 \text{ for each } j. \tag{Assumption 6}$$

Informally, we need to ensure that diffraction patterns of strained crystals still look like blurred diffraction patterns of single crystals. That is to say that the diffraction patterns should still be essentially sharp with isolated, if blurred, Bragg disks.

## 3.4 Linearised model of electron diffraction from deformed crystals

A linear tomography model is developed from the kinematical diffraction model in this section following a physically motivated argument based on precession electron diffraction, which is supported by a parallel mathematically rigorous argument in Section 3.5. It is not clear how the necessary assumptions may be justified physically and we therefore provide computational results in Section 3.7 which demonstrate that our model can be quantitatively accurate.

### 3.4.1 Strained diffraction patterns with precession

In Section 1.4.5, precession was motivated as a technique for simplifying the computation of the average deformation. Now, we make this statement more precise by using tools from probability theory.

**Approximation 3.1.** *If the beam energy is large and the strain is small (Assumption 2 and Assumption 6) then*

$$D_\alpha(\boldsymbol{k}) = \overline{D}_\alpha(\boldsymbol{k}) + \mathit{error} \tag{3.5}$$

$$\text{where} \quad \overline{D}_\alpha(\boldsymbol{k}) := \left|\sum_{i=1}^\infty \overline{a}_i \mathop{\mathbb{E}}_{\Gamma^\top\beta_j=0} \mathrm{f}(\boldsymbol{k} - \Gamma^\top A_j^\top \boldsymbol{p}_i)\right|^2 \tag{3.6}$$

*for some new weights $\overline{a}_i \in \mathbb{C}$.*

$\overline{D}_\alpha$ is exactly the simple model one would hope for in strain mapping. Ignoring the squared norm, such a diffraction pattern is the average of each idealised diffraction pattern with an average spot-shape and weighted with an average structure-factor $\overline{a}_i$ independent of the strain parameter $A_j$. The cost for assuming that the raw data $D_\alpha$ obeys such a simple model is determined by the error term. With too little precession $D_\alpha$ will look close to Figure 3.1.a with inhomogeneous spot intensities and so the error will be large, however, too much precession introduces a new blurring of the data which also contributes to the error. The important point is that the error should not bias the computations that we go on to make in Approximation 3.2, in particular it should not bias the centres of spots away from $\overline{D}_\alpha$.

Algebraically, it is hard to quantify the precise sweet-spot but the proof motivates a rule-of-thumb for the choice of precession angle. It should be just sufficiently large to ensure that

$$k_z(\boldsymbol{k}) - (R_t^\top A_j^\top \boldsymbol{p}_i)_z = 0 \qquad \text{for some } t \in [0, 2\pi),\ |\boldsymbol{k}| = |\boldsymbol{p}_i|$$

for all $j$ and $i$ such that $p_{i,z} = 0$. This is sufficient to guarantee that for any spot visible in Figure 3.1 and all small strains ($|A_j - \mathrm{id}\,| < \sigma$), the precession angle $\alpha$ is large enough such that $R_t^\top A_j^\top \boldsymbol{p}_i$ lies on the Ewald sphere for some $t$. After a geometrical argument detailed in Section B.3, to analyse the deformation of spots $\boldsymbol{p}_i$ with $|\boldsymbol{p}_i| < P$, we suggest the relation

$$\alpha \approx \cos^{-1}\left(1 - \frac{\sigma^2}{2}\right) + \sin^{-1}\left(\frac{\lambda P}{4\pi}\right) \sim \sigma + \frac{\lambda P}{4\pi}. \tag{3.7}$$

In words, the precession angle should be larger than the maximum rotation due to the deformation plus the distance between the flat hyperplane and the Ewald sphere.

One final observation from this approximation is that the coordinate $(A_j^\top \boldsymbol{p}_i)_z$ has disappeared completely, other than in the choice of $\alpha$ above, and the remaining expression only depends on $\Gamma^\top A_j^\top \boldsymbol{p}_i$. This indicates that diffraction imaging is insensitive to deformations parallel to the beam direction, which will dictate our choice of tensor tomography model in Section 3.6.

### 3.4.2 Linearised diffraction model

The final approximation is to linearise the forward model. $\overline{D}_\alpha$ is now a sufficiently simple, but still non-linear, model to go from a deformed crystal to a diffraction image, however, what we want is a simple linear model to map from a deformed crystal to its average deformation tensor. To do this we propose a simple pre-processing procedure to apply to the raw data $D_\alpha$ which corresponds to a linear forward model with respect to the deformation parameters. In particular, we propose computing centres of mass for each of the observed diffraction spots.

**Approximation 3.2.** *If the conditions of Approximation 3.1 hold and the diffracted spots are symmetric and non-overlapping (Assumption 3 and Assumption 4) then*

$$\mathop{\mathbb{E}}_{\Gamma^\top \beta_j = 0} \Gamma^\top A_j^\top \boldsymbol{p}_i = \frac{\int_{|\boldsymbol{k}-\boldsymbol{q}_i|<\overline{r}} \boldsymbol{k} D_\alpha(\boldsymbol{k}) d\boldsymbol{k}}{\int_{|\boldsymbol{k}-\boldsymbol{q}_i|<\overline{r}} D_\alpha(\boldsymbol{k}) d\boldsymbol{k}} + \mathit{error} \tag{3.8}$$

*where $\overline{r} > 0$ is the separation of spots given by Assumption 3.*

This theorem directly indicates that centres of mass are a good linear model for deformed diffraction patterns and this is confirmed by the numerical results in Section 3.7. More generally, this provides a motivation that the centre of deformed spots are equal to average deformation tensors. Other centre detection methods are common in the literature and we also give numerical comparison of the accuracy and robustness of each method.

## 3.5 Analytical justification of Section 3.4

In Section 3.4 we give a physical motivation for why centre of mass calculations should accurately predict average strain values. In this section we provide a formal mathematical link from the precessed Ewald sphere model to this same goal. Some of the statistical assumptions made here are highly technical and it would be hard to justify them from a purely physical perspective. However, as in the main text, the core justification which we rely upon is the simulation study of the whole pipeline.

In the remainder of this section, we first sketch the proof at a high level listing a sequence of results and then the details of the longer proofs appear at the end of the section.

To recap, the starting point of this argument is the physical model which we assume to be exact. In particular,

$$D(\boldsymbol{k}) = |\mathcal{F}[\Psi_p] \star \mathcal{F}[u]|^2(\boldsymbol{k}, k_z(\boldsymbol{k}))$$

where

$$u(\boldsymbol{r}) = \sum_{j=1}^{N} u_0(A_j\boldsymbol{r} + \boldsymbol{b}_j)\mathbb{1}_{|\cdot|_\infty \leq 1/2}\left(\frac{\boldsymbol{r} - \boldsymbol{\beta}_j}{\rho}\right).$$

The first step is to spatially localise the strain which generates the signal in a diffraction pattern. A key part of tensor tomography is that signal only depends on the strain contained in blocks (or voxels) which directly intersect the electron beam. This is formalised in the following result which also expands the notation to take advantage of the particular structure of $u$.

**Theorem** (Lemma 3.3.4)**.** *If Assumption 5 (narrow probe) holds then*

$$D(\boldsymbol{k}) = \left| \sum_{i\in\mathbb{N}, j\in[N]} \widehat{a}_i(k_z(\boldsymbol{k}) - (A_j^\top \boldsymbol{p}_i)_z, \beta_{j,z})e^{\imath \boldsymbol{b}_j \bullet \boldsymbol{p}_i}\, \mathrm{f}(\boldsymbol{k} - \Gamma^\top A_j^\top \boldsymbol{p}_i) \right|^2$$

$$\textit{where}\ \ \mathrm{f}(\boldsymbol{k}) = \left[\mathcal{F}[\Psi_p] \star [\rho^2 \operatorname{sinc}(\rho\boldsymbol{k}')]\right](\boldsymbol{k})$$

$$\textit{and}\ \widehat{a}_i(k, \beta) = a_i\rho \operatorname{sinc}(\rho k)e^{-\imath\beta k}.$$

As can be seen, this expression is still highly non-linear in terms of the strain parameter $A_j$ and indeed too non-linear for us to develop a linear model directly. To make this possible we introduce the beam precession technique. The heuristic is that if we modify our data at acquisition then we can analyse it as if it were generated by a simpler model. There is of course a trade-off here, a small amount precession will usefully smooth the problem but too much will begin to blur out the desired structures. The following two lemmas make this statement precise and their combination is exactly the statement of Approximation 3.1.

**Lemma 3.5.1.**

$$D_\alpha(\boldsymbol{k}) = \left|\mathop{\mathbb{E}}_t \mathcal{F}[\Psi_p u(R_t\boldsymbol{r})](\boldsymbol{k}, k_z(\boldsymbol{k}))\right|^2 + \mathrm{Var}_t\left[\mathcal{F}[\Psi_p u(R_t\boldsymbol{r})](\boldsymbol{k}, k_z(\boldsymbol{k}))\right]$$

*Proof.* This is just the definition of variance,

$$\mathop{\mathbb{E}}_t |Y_t|^2 = |\mathop{\mathbb{E}}_t Y_t|^2 + \mathrm{Var}_t\, Y_t$$

for the appropriate choice of random variable $Y_t$. □

This claim is the most physically ambiguous. It suggests that coherent precession is a good approximation of incoherent precession. The justification of this is that it only needs to be true when looking at the whole pipeline. In particular, the variance term may be large in magnitude but so long as it does not strongly bias the locations of the centres of the diffracted peaks then it will not affect the overall approximation.

**Lemma 3.5.2.** *If the $A_jR_t\Gamma$, $A_jR_t\left(\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right)$, and $\boldsymbol{b}_j$ are independent random variables over $(j,t)$ and the high-energy limit is valid (Assumption 2) then*

$$\left|\mathop{\mathbb{E}}_t \mathcal{F}[\Psi_p u(R_t\boldsymbol{r})](\boldsymbol{k}, k_z(\boldsymbol{k}))\right|^2 = \underbrace{\left|\sum_{i=1}^{\infty} \overline{a}_i \mathop{\mathbb{E}}_{\Gamma^\top\boldsymbol{\beta}_j=0} \mathrm{f}(\boldsymbol{k} - \Gamma^\top A_j^\top \boldsymbol{p}_i)\right|^2}_{=:\overline{D}_\alpha} + O(\alpha^2)$$

*for some new weights $\overline{a}_i \in \mathbb{C}$.*

While the assumptions of this lemma appear highly technical they are also not unreasonable from a physical standpoint. $\boldsymbol{b}_j$ and $A_j$ represent translations and strains respectively. These are physically distinct quantities and so $\boldsymbol{b}_j$ should also be statistically independent from the first two terms. Looking closer at these terms, $A_jR_t\Gamma$ is the first two columns of $A_jR_t$ and $A_jR_t\left(\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right)$ is the third. If $\alpha$ is small then this is just a statement that the out-of-plane strain is independent to the in-plane strain.

Finally, we need to justify that the centre of mass is an accurate predictor of average strain. The difficulty here again is the squared modulus, if the exit waves were imaged directly then this step would be direct but the square has the potential to bias towards points of maximum intensity. The assumptions necessary at this stage are some form of symmetry on the strain field, the technical statement is as follows.

**Lemma 3.5.3.** *Suppose the conditions of Approximation 3.1 hold and the diffracted spots are symmetric and non-overlapping (Assumption 3 and Assumption 4). If the random variables $A_{j_1} + A_{j_2}$ and $A_{j_1} - A_{j_2}$ are independent over random pairs of indices $\{(j_1, j_2)$ s.t. $\Gamma^\top\boldsymbol{\beta}_{j_1} =$*

$\Gamma^\top \boldsymbol{\beta}_{j_2} = \mathbf{0}\}$ *then for each $i$:*

$$\mathop{\mathbb{E}}_{\Gamma^\top \boldsymbol{\beta}_j = 0} \Gamma^\top A_j^\top \boldsymbol{p}_i = \frac{\int_{|\boldsymbol{k} - \Gamma^\top \boldsymbol{p}_i| < \overline{r}} \boldsymbol{k} \overline{D}_\alpha(\boldsymbol{k}) d\boldsymbol{k}}{\int_{|\boldsymbol{k} - \Gamma^\top \boldsymbol{p}_i| < \overline{r}} \overline{D}_\alpha(\boldsymbol{k}) d\boldsymbol{k}}$$

*whenever the denominator is not 0 and where $\overline{r} > 0$ is the separation of spots given by Assumption 3.*

Approximation 3.2 is the combination of the previous three lemmas where any violations of the exact conditions becomes absorbed into the error term. We now provide the proofs of these lemmas.

**Lemma 3.5.4.** *For all $t \in [0, 2\pi]$*

$$R_t\Gamma - \mathop{\mathbb{E}}_t R_t\Gamma = O(\alpha).$$

*Proof.* This is a purely algebraic proof:

$$\begin{aligned}
R_t\Gamma &= \begin{pmatrix} \cos(t) & \sin(t) & 0 \\ -\sin(t) & \cos(t) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) \\ 0 & -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} \cos(t) & -\sin(t) & 0 \\ \sin(t) & \cos(t) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \\
&= \begin{pmatrix} \cos(t) & \sin(t) & 0 \\ -\sin(t) & \cos(t) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) \\ 0 & -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \\ 0 & 0 \end{pmatrix} \\
&= \begin{pmatrix} \cos(t) & \sin(t) & 0 \\ -\sin(t) & \cos(t) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(t) & -\sin(t) \\ \cos(\alpha)\sin(t) & \cos(\alpha)\cos(t) \\ -\sin(\alpha)\sin(t) & -\sin(\alpha)\cos(t) \end{pmatrix} \\
&= \begin{pmatrix} \cos^2(t) + \cos(\alpha)\sin^2(t) & (\cos(\alpha) - 1)\cos(t)\sin(t) \\ (\cos(\alpha) - 1)\cos(t)\sin(t) & \sin^2(t) + \cos(\alpha)\cos^2(t) \\ -\sin(\alpha)\sin(t) & -\sin(\alpha)\cos(t) \end{pmatrix} \\
&= \Gamma + \begin{pmatrix} (\cos(\alpha) - 1)\sin^2(t) & (\cos(\alpha) - 1)\cos(t)\sin(t) \\ (\cos(\alpha) - 1)\cos(t)\sin(t) & (\cos(\alpha) - 1)\cos^2(t) \\ -\sin(\alpha)\sin(t) & -\sin(\alpha)\cos(t) \end{pmatrix} \\
&= \Gamma - \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ \sin(t) & \cos(t) \end{pmatrix} \sin\alpha + O(\alpha^2).
\end{aligned}$$

This gives

$$\begin{aligned}
R_t\Gamma - \mathop{\mathbb{E}}_t R_t\Gamma &= -\left[ \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ \sin(t) & \cos(t) \end{pmatrix} - \mathop{\mathbb{E}}_t \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ \sin(t) & \cos(t) \end{pmatrix} \right] \sin\alpha + O(\alpha^2) \\
&= O(\alpha).
\end{aligned}$$

□

*Proof of Lemma 3.5.2.* Note that we can re-write

$$\begin{aligned}
&\left|\mathop{\mathbb{E}}_{t} \mathcal{F}[\Psi_p u(R_t \boldsymbol{r})](\boldsymbol{k}, k_z(\boldsymbol{k}))\right|^2 \\
&\qquad\qquad = \left| \sum_{i\in\mathbb{N}, \Gamma^\top \boldsymbol{\beta}_j = 0} \mathop{\mathbb{E}}_{t} \left[ \widehat{a}_i(k_z(\boldsymbol{k}) - ((A_j R_t)^\top \boldsymbol{p}_i)_z, \beta_{j,z}) e^{\imath \boldsymbol{b}_j \bullet \boldsymbol{p}_i} \,\mathrm{f}(\boldsymbol{k} - \Gamma^\top (A_j R_t)^\top \boldsymbol{p}_i) \right] \right|^2 \\
&\qquad\qquad = \left| \sum_{i\in\mathbb{N}} \mathop{\mathbb{E}}_{j,t} \left[ X_{i,j,t}(\boldsymbol{k}) Y_{i,j,t}(\boldsymbol{k}) Z_{i,j,t}(\boldsymbol{k}) \right] \right|^2 \times |\{j \text{ s.t. } \Gamma^\top \boldsymbol{\beta}_j = \boldsymbol{0}\}|^2
\end{aligned}$$

where the constant comes from switching the sum to average over $j$ and we define

$$\begin{aligned}
X_{i,j,t}(\boldsymbol{k}) &= \widehat{a}_i(k_z(\boldsymbol{k}) - (R_t^\top A_j^\top \boldsymbol{p}_i)_z, \beta_{j,z}) && = X_i\left(\boldsymbol{k} | A_j R_t \left(\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right)\right), \\
Y_{i,j,t}(\boldsymbol{k}) &= \exp(\imath \boldsymbol{b}_j \bullet \boldsymbol{p}_i) && = Y_i(\boldsymbol{k} | \boldsymbol{b}_j), \\
Z_{i,j,t}(\boldsymbol{k}) &= \mathrm{f}(\boldsymbol{k} - \Gamma^\top R_t^\top A_j^\top \boldsymbol{p}_i) && = Z_i(\boldsymbol{k} | A_j R_t \Gamma).
\end{aligned}$$

These functions should be considered as a set of random variables which, for each $\boldsymbol{k}$, are indexed over $i$ and the indices $(j,t)$ are considered the source of randomness. Restating the assumptions of the theorem onto these variables, we know $X_i$, $Y_i$, and $Z_i$ are independent in $(j,t)$. With this, we can apply Lemma B.2.2 to simplify

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{t} \mathcal{F}[\Psi_p u(R_t \boldsymbol{r})](\boldsymbol{k}, k_z(\boldsymbol{k}))\right|^2 &\propto \left| \sum_{i\in\mathbb{N}} \mathop{\mathbb{E}}_{j,t} \left[ X_{i,j,t}(\boldsymbol{k}) Y_{i,j,t}(\boldsymbol{k}) Z_{i,j,t}(\boldsymbol{k}) \right] \right|^2 \\
&= \left| \sum_{i\in\mathbb{N}} \mathbb{E}\left[X_i(\boldsymbol{k})\right] \mathbb{E}\left[Y_i(\boldsymbol{k})\right] \mathbb{E}\left[Z_i(\boldsymbol{k})\right] \right|^2
\end{aligned}$$

Each of these factors now simplifies:

- Assuming the high-energy limit we have $k_z(\boldsymbol{k}) = 0$ and so $\mathbb{E}\, X_i(\boldsymbol{k}) = \text{constant}_i$.

- $Y_i$ is not a function of $\boldsymbol{k}$ so we trivially have $\mathbb{E}\, Y_i(\boldsymbol{k}) = \text{constant}'_i$.

- By Lemma B.2.3 we can approximate

$$\begin{aligned}
\mathbb{E}\, Z_i(\boldsymbol{k}) = \mathop{\mathbb{E}}_{j} \mathop{\mathbb{E}}_{t} \mathrm{f}(\boldsymbol{k} - \Gamma^\top R_t^\top A_j^\top \boldsymbol{p}_i) &= \mathop{\mathbb{E}}_{j} \mathrm{f}(\boldsymbol{k} - \mathop{\mathbb{E}}_{t} \Gamma^\top R_t^\top A_j^\top \boldsymbol{p}_i) + O(|R_t \Gamma - \mathop{\mathbb{E}}_{t} R_t \Gamma|^2) \\
&= \mathop{\mathbb{E}}_{j} \mathrm{f}(\boldsymbol{k} - \mathop{\mathbb{E}}_{t} \Gamma^\top R_t^\top A_j^\top \boldsymbol{p}_i) + O(\alpha^2)
\end{aligned}$$

The theorem is concluded by gathering the new constants into $\overline{a}_i$. □

**Lemma 3.5.5.** *Under Assumption 4,*

$$\int_{\mathbb{R}^2} \boldsymbol{k}\, \mathrm{f}(\boldsymbol{k}-\boldsymbol{c})\, \mathrm{f}(\boldsymbol{k}+\boldsymbol{c}) d\boldsymbol{k} = \boldsymbol{0}$$

*for all* $\boldsymbol{c}$*.*

*Proof.* The proof is direct:

$$\begin{aligned} \int_{\mathbb{R}^2} \boldsymbol{k}\, \mathrm{f}(\boldsymbol{k}-\boldsymbol{c})\, \mathrm{f}(\boldsymbol{k}+\boldsymbol{c}) d\boldsymbol{k} &= \int_{\mathbb{R}^2} -\boldsymbol{k}\, \mathrm{f}(-\boldsymbol{k}-\boldsymbol{c})\, \mathrm{f}(-\boldsymbol{k}+\boldsymbol{c}) |\det(-\mathrm{id})| d\boldsymbol{k} \\ &= -\int_{\mathbb{R}^2} \boldsymbol{k}\, \mathrm{f}(\boldsymbol{k}+\boldsymbol{c})\, \mathrm{f}(\boldsymbol{k}-\boldsymbol{c}) d\boldsymbol{k}. \qquad \text{Assumption 4} \end{aligned}$$

□

*Proof of Lemma 3.5.3.* To compute centres of mass, it shall be convenient to define some new functions and abbreviations. We define the functions:

$$\mathrm{E}_0(\boldsymbol{c}) = \int_{\mathbb{R}^2} \mathrm{f}(\boldsymbol{k} - {}^1\!/_2\boldsymbol{c})\, \mathrm{f}(\boldsymbol{k} + {}^1\!/_2\boldsymbol{c}) d\boldsymbol{k}, \qquad \mathrm{E}_1(\boldsymbol{c}) = \int_{\mathbb{R}^2} \boldsymbol{k}\, \mathrm{f}(\boldsymbol{k} - {}^1\!/_2\boldsymbol{c})\, \mathrm{f}(\boldsymbol{k} + {}^1\!/_2\boldsymbol{c}) d\boldsymbol{k}$$

and the points

$$\boldsymbol{q}_i = \Gamma^\top \boldsymbol{p}_i, \qquad \boldsymbol{q}_{i,j} = \Gamma^\top A_j^\top \boldsymbol{p}_i$$

to remove excessive use of the $\Gamma^\top$ projection. Also note by Lemma 3.5.5 that $\mathrm{E}_1(\boldsymbol{c}) = 0$ for all $\boldsymbol{c}$. We now compute the relevant integrals starting with the formula of (3.6):

$$\int_{|\boldsymbol{k}-\boldsymbol{q}_i|<\overline{r}} \overline{D}_\alpha(\boldsymbol{k}) d\boldsymbol{k} = \int_{|\boldsymbol{k}-\boldsymbol{q}_i|<\overline{r}} \left| \sum_{i'\in\mathbb{N}} \underset{\Gamma^\top \boldsymbol{\beta}_j = \boldsymbol{0}}{\mathbb{E}} \left[ \overline{a}_{i'}\, \mathrm{f}(\boldsymbol{k} - \Gamma^\top \boldsymbol{q}_{i',j}) \right] \right|^2 d\boldsymbol{k}.$$

Using Assumption 3, we know that the only $i'$ for which the summand is non-zero on this integral domain is $i' = i$. Also using Assumption 3, we know that $\mathrm{f}(\boldsymbol{k} - \boldsymbol{q}_{i,j}) = 0$ outside of the restricted domain and so we drop this constraint to simplify notation.

$$\int_{|\boldsymbol{k}-\boldsymbol{q}_i|<\overline{r}} \overline{D}_\alpha(\boldsymbol{k}) d\boldsymbol{k} = \int_{|\boldsymbol{k}-\boldsymbol{q}_i|<\overline{r}} \left| \underset{\Gamma^\top \boldsymbol{\beta}_j = \boldsymbol{0}}{\mathbb{E}} \left[ \overline{a}_i\, \mathrm{f}(\boldsymbol{k} - \boldsymbol{q}_{i,j}) \right] \right|^2 d\boldsymbol{k} = \int_{\mathbb{R}^2} \left| \underset{\Gamma^\top \boldsymbol{\beta}_j = \boldsymbol{0}}{\mathbb{E}} \left[ \overline{a}_i\, \mathrm{f}(\boldsymbol{k} - \boldsymbol{q}_{i,j}) \right] \right|^2 d\boldsymbol{k}.$$

Next, we expand the brackets noting that $f$ is a real valued function:

$$\begin{aligned} \int_{|\boldsymbol{k}-\boldsymbol{q}_i|<\overline{r}} \overline{D}_\alpha(\boldsymbol{k}) d\boldsymbol{k} &= \int_{\mathbb{R}^2} |\overline{a}_i|^2 \underset{\substack{\Gamma^\top \boldsymbol{\beta}_j = \boldsymbol{0} \\ \Gamma^\top \boldsymbol{\beta}_J = \boldsymbol{0}}}{\mathbb{E}} \left[ \mathrm{f}(\boldsymbol{k} - \boldsymbol{q}_{i,j})\, \mathrm{f}(\boldsymbol{k} - \Gamma^\top \boldsymbol{q}_{i,J}) d\boldsymbol{k} \right] \\ &= |\overline{a}_i|^2 \underset{\substack{\Gamma^\top \boldsymbol{\beta}_j = \boldsymbol{0} \\ \Gamma^\top \boldsymbol{\beta}_J = \boldsymbol{0}}}{\mathbb{E}} \left[ \int_{\mathbb{R}^2} \mathrm{f}(\boldsymbol{k} - \boldsymbol{q}_{i,j})\, \mathrm{f}(\boldsymbol{k} - \Gamma^\top \boldsymbol{q}_{i,J}) d\boldsymbol{k} \right]. \end{aligned}$$

The coordinate translation $\boldsymbol{k} \mapsto \boldsymbol{k} - \frac{1}{2}(\boldsymbol{q}_{i,j} + \boldsymbol{q}_{i,J})$ simplifies this to a special case of $\mathrm{E}_0$:

$$\begin{aligned}\int_{|\boldsymbol{k}-\boldsymbol{q}_i|<\overline{r}} \overline{D}_\alpha(\boldsymbol{k}) d\boldsymbol{k} &= |\overline{a}_i|^2 \mathop{\mathbb{E}}_{\substack{\Gamma^\top \boldsymbol{\beta}_j=\boldsymbol{0}\\ \Gamma^\top \boldsymbol{\beta}_J=\boldsymbol{0}}} \left[ \int_{\mathbb{R}^2} \mathrm{f}(\boldsymbol{k} - {}^1\!/\!_2(\boldsymbol{q}_{i,j} - \boldsymbol{q}_{i,J}))\, \mathrm{f}(\boldsymbol{k} + {}^1\!/\!_2(\boldsymbol{q}_{i,j} - \boldsymbol{q}_{i,J})) d\boldsymbol{k} \right] \\ &= |\overline{a}_i|^2 \mathop{\mathbb{E}}_{\substack{\Gamma^\top \boldsymbol{\beta}_j=\boldsymbol{0}\\ \Gamma^\top \boldsymbol{\beta}_J=\boldsymbol{0}}} \mathrm{E}_0(\boldsymbol{q}_{i,j} - \boldsymbol{q}_{i,J}).\end{aligned}$$

Similarly,

$$\begin{aligned}\int_{|\boldsymbol{k}-\boldsymbol{q}_i|<\overline{r}} \boldsymbol{k}\overline{D}_\alpha(\boldsymbol{k}) d\boldsymbol{k} &= \int_{|\boldsymbol{k}-\boldsymbol{q}_i|<\overline{r}} \boldsymbol{k} \left| \sum_{i' \in \mathbb{N}} \mathop{\mathbb{E}}_{\Gamma^\top \boldsymbol{\beta}_j=\boldsymbol{0}} \left[ \overline{a}_{i'}\, \mathrm{f}(\boldsymbol{k} - \boldsymbol{q}_{i',j}) \right] \right|^2 d\boldsymbol{k} \\ &= |\overline{a}_i|^2 \mathop{\mathbb{E}}_{\substack{\Gamma^\top \boldsymbol{\beta}_j=\boldsymbol{0}\\ \Gamma^\top \boldsymbol{\beta}_J=\boldsymbol{0}}} \left[ \int_{\mathbb{R}^2} \boldsymbol{k}\, \mathrm{f}(\boldsymbol{k} - \boldsymbol{q}_{i,j})\, \mathrm{f}(\boldsymbol{k} - \boldsymbol{q}_{i,J}) d\boldsymbol{k} \right] \\ &= |\overline{a}_i|^2 \mathop{\mathbb{E}}_{\substack{\Gamma^\top \boldsymbol{\beta}_j=\boldsymbol{0}\\ \Gamma^\top \boldsymbol{\beta}_J=\boldsymbol{0}}} \left[ \int_{\mathbb{R}^2} \left( \boldsymbol{k} + \frac{\boldsymbol{q}_{i,j} + \boldsymbol{q}_{i,J}}{2} \right) f \left( \boldsymbol{k} - \frac{\boldsymbol{q}_{i,j} - \boldsymbol{q}_{i,J}}{2} \right) f \left( \boldsymbol{k} + \frac{\boldsymbol{q}_{i,j} - \boldsymbol{q}_{i,J}}{2} \right) d\boldsymbol{k} \right] \\ &= |\overline{a}_i|^2 \mathop{\mathbb{E}}_{\substack{\Gamma^\top \boldsymbol{\beta}_j=\boldsymbol{0}\\ \Gamma^\top \boldsymbol{\beta}_J=\boldsymbol{0}}} \left[ \mathrm{E}_1(\boldsymbol{q}_{i,j} - \boldsymbol{q}_{i,J}) + \frac{\boldsymbol{q}_{i,j} + \boldsymbol{q}_{i,J}}{2} \mathrm{E}_0(\boldsymbol{q}_{i,j} - \boldsymbol{q}_{i,J}) \right] \\ &= |\overline{a}_i|^2 \mathop{\mathbb{E}}_{\substack{\Gamma^\top \boldsymbol{\beta}_j=\boldsymbol{0}\\ \Gamma^\top \boldsymbol{\beta}_J=\boldsymbol{0}}} \left[ \frac{\boldsymbol{q}_{i,j} + \boldsymbol{q}_{i,J}}{2} \mathrm{E}_0(\boldsymbol{q}_{i,j} - \boldsymbol{q}_{i,J}) \right].\end{aligned}$$

Finally, as $A_j + A_J$ is independent of $A_j - A_J$ we can translate this to $\boldsymbol{q}_{i,j}/\boldsymbol{q}_{i,J}$ and again apply Lemma B.2.2:

$$\mathop{\mathbb{E}}_{j,J} \left[ \frac{\boldsymbol{q}_{i,j} + \boldsymbol{q}_{i,J}}{2} \mathrm{E}_0(\boldsymbol{q}_{i,j} - \boldsymbol{q}_{i,J}) \right] = \mathop{\mathbb{E}}_{j,J} \left[ \frac{\boldsymbol{q}_{i,j} + \boldsymbol{q}_{i,J}}{2} \right] \mathop{\mathbb{E}}_{j,J} \left[ \mathrm{E}_0(\boldsymbol{q}_{i,j} - \boldsymbol{q}_{i,J}) \right]. \tag{3.9}$$

Thus

$$\begin{aligned}\frac{\int_{|\boldsymbol{k}-\boldsymbol{q}_i|<\overline{r}} \boldsymbol{k}\overline{D}_\alpha(\boldsymbol{k}) d\boldsymbol{k}}{\int_{|\boldsymbol{k}-\boldsymbol{q}_i|<\overline{r}} \overline{D}_\alpha(\boldsymbol{k}) d\boldsymbol{k}} &= \frac{|\overline{a}_i|^2\, \mathbb{E}_{j,J} \left[ \frac{\boldsymbol{q}_{i,j}+\boldsymbol{q}_{i,J}}{2} \right] \mathbb{E}_{j,J} \left[ \mathrm{E}_0(\boldsymbol{q}_{i,j} - \boldsymbol{q}_{i,J}) \right]}{|\overline{a}_i|^2\, \mathbb{E}_{j,J} \left[ \mathrm{E}_0(\boldsymbol{q}_{i,j} - \boldsymbol{q}_{i,J}) \right]} \\ &= \frac{1}{2} \mathop{\mathbb{E}}_{j,J} \left[ \boldsymbol{q}_{i,j} + \boldsymbol{q}_{i,J} \right] = \mathop{\mathbb{E}}_{\Gamma^\top \boldsymbol{\beta}_j=0} \boldsymbol{q}_{i,j} = \mathop{\mathbb{E}}_{\Gamma^\top \boldsymbol{\beta}_j=0} \Gamma^\top A_j^\top \boldsymbol{p}_i.\end{aligned}$$

□

## 3.6 Non-symmetric tensor tomography

Up to this point we have considered how to compute a single average deformation tensor from a single diffraction pattern. The role of this section is to identify this process with an inverse problem capable of reconstructing a 3D strain map from many average deformation tensors and then highlight the relevant properties of this inverse problem.

The fact that strain maps are rank-2 tensor fields immediately puts us in the domain of tensor tomography and the physical characteristics of diffraction imaging seen in Approximation 3.1 highlights the transverse ray transform (TRT) as the natural parallel. In particular, Approximation 3.1 shows that precessed diffraction patterns are insensitive to out-of-plane strain. This aligns with equivalent reasoning by Lionheart and Withers (2015) for polycrystalline materials where it was shown that this corresponds to the TRT. The only difference for polycrystalline materials is that the forward model is insensitive to the skew component of deformations and so we will extend the analysis of the TRT to account for general tensor fields of non-symmetric tensors.

### 3.6.1 Electron diffraction and the transverse ray transform

The first step is to generalise the notation from Approximation 3.2 to consider diffraction patterns where the electron beam is not parallel to the $z$-axis or through the point $x = y = 0$. This is mainly an algebraic exercise, first to upgrade from average spot centres to average strain tensors and then to realise the generalisation.

**Lemma 3.6.1.**

1. Tensors from vectors: $\mathbb{E}_{\Gamma^\top\beta_j=0}\Gamma^\top A_j^\top\Gamma$ *can be computed from the values of* $\mathbb{E}_{\Gamma^\top\beta_j=0}\Gamma^\top A_j^\top\boldsymbol{p}_i$ *for any two non-colinear points, say* $\boldsymbol{p}_1, \boldsymbol{p}_2$*, such that* $p_{i,z} = 0$.

2. Generalising notation: *If we choose* $\boldsymbol{r} = \left(\begin{smallmatrix}0\\0\\0\end{smallmatrix}\right)$ *and* $\Pi_{\boldsymbol{\theta}} = \mathrm{id} - \frac{\boldsymbol{\theta}\boldsymbol{\theta}^\top}{|\boldsymbol{\theta}|^2}$ *with* $\boldsymbol{\theta} = \left(\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right)$ *then*

$$\underset{\Gamma^\top\beta_j=\boldsymbol{r}}{\mathbb{E}} \Pi_{\boldsymbol{\theta}}^\top A_j^\top \Pi_{\boldsymbol{\theta}} = \begin{pmatrix} \underset{\Gamma^\top\beta_j=\boldsymbol{0}}{\mathbb{E}} \Gamma^\top A_j^\top \Gamma & \begin{matrix}0\\0\end{matrix} \\ \begin{matrix}0 & 0\end{matrix} & 0 \end{pmatrix}.$$

*Proof.* Let square brackets temporarily denote the horizontal concatenation of vectors into matrices. Note that

$$[\boldsymbol{p}_1, \boldsymbol{p}_2] = \left(\begin{smallmatrix} p_{1,x} & p_{2,x} \\ p_{1,y} & p_{2,y} \\ 0 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} p_{1,x} & p_{2,x} \\ p_{1,y} & p_{2,y} \end{smallmatrix}\right) = \Gamma\left(\begin{smallmatrix} p_{1,x} & p_{2,x} \\ p_{1,y} & p_{2,y} \end{smallmatrix}\right).$$

Vectors $\boldsymbol{p}_i$ are non-co-linear and so this final matrix is invertible. Thus, we have

$$\begin{aligned}
\underset{\Gamma^\top\boldsymbol{\beta}_j=0}{\mathbb{E}} \Gamma^\top A_j^\top \Gamma &= \underset{\Gamma^\top\boldsymbol{\beta}_j=0}{\mathbb{E}} \left\{\Gamma^\top A_j^\top \Gamma \left(\begin{smallmatrix} p_{1,x} & p_{2,x} \\ p_{1,y} & p_{2,y}\end{smallmatrix}\right)\left(\begin{smallmatrix} p_{1,x} & p_{2,x} \\ p_{1,y} & p_{2,y}\end{smallmatrix}\right)^{-1}\right\} \\
&= \underset{\Gamma^\top\boldsymbol{\beta}_j=0}{\mathbb{E}} \left\{\Gamma^\top A_j^\top [\boldsymbol{p}_1, \boldsymbol{p}_2]\right\}\left(\begin{smallmatrix} p_{1,x} & p_{2,x} \\ p_{1,y} & p_{2,y}\end{smallmatrix}\right)^{-1} \\
&= \left[\underset{\Gamma^\top\boldsymbol{\beta}_j=0}{\mathbb{E}} \Gamma^\top A_j^\top \boldsymbol{p}_1, \underset{\Gamma^\top\boldsymbol{\beta}_j=0}{\mathbb{E}} \Gamma^\top A_j^\top \boldsymbol{p}_2\right]\left(\begin{smallmatrix} p_{1,x} & p_{2,x} \\ p_{1,y} & p_{2,y}\end{smallmatrix}\right)^{-1}
\end{aligned}$$

which verifies the first part. The final part is a simple algebraic argument:

$$\Pi_{\boldsymbol{\theta}} = \mathrm{id} - \left(\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right)\left(\begin{smallmatrix}0&0&1\end{smallmatrix}\right) = \left(\begin{smallmatrix}1&0&0\\0&1&0\\0&0&0\end{smallmatrix}\right) = \left(\begin{smallmatrix}1&0\\0&1\\0&0\end{smallmatrix}\right)\left(\begin{smallmatrix}1&0&0\\0&1&0\end{smallmatrix}\right) = \Gamma\Gamma^\top.$$

From this we see

$$\underset{\Gamma^\top\boldsymbol{\beta}_j=\boldsymbol{r}}{\mathbb{E}} \Pi_{\boldsymbol{\theta}}^\top A_j^\top \Pi_{\boldsymbol{\theta}} = \Gamma\left\{\underset{\Gamma^\top\boldsymbol{\beta}_j=0}{\mathbb{E}} \Gamma^\top A_j \Gamma\right\}\Gamma^\top = \begin{pmatrix} \mathbb{E}_{\Gamma^\top\boldsymbol{\beta}_j=0}\Gamma^\top A_j^\top \Gamma & & 0 \\ & & 0 \\ 0 & 0 & 0\end{pmatrix}$$

as required. □

This lemma is where Assumption 1 becomes relevant. If the crystal is of finite thickness then all $\boldsymbol{p}_i$ are always present in the diffraction pattern (not just when $p_{i,z} = 0$) although the intensity may be small. In practice, all visible spots will lie on this hyperplane which is justified by Assumption 1.

The final step to aligning with the TRT is to replace discrete sums with line integrals. We compute

$$\begin{aligned}
\underset{\Gamma^\top\boldsymbol{\beta}_j=\boldsymbol{r}}{\mathbb{E}} \Pi_{\boldsymbol{\theta}}^\top A_j^\top \Pi_{\boldsymbol{\theta}} &= \frac{\int_{\Gamma^\top\boldsymbol{\beta}_j=\boldsymbol{0}} \Pi_{\boldsymbol{\theta}}^\top A(\boldsymbol{\beta}_j)^\top \Pi_{\boldsymbol{\theta}}\, dj}{|\{j \text{ s.t. } \Gamma^\top\boldsymbol{\beta}_j = \boldsymbol{0}\}|} && (3.10) \\
&\propto \int_{\mathbb{R}} \Pi_{\boldsymbol{\theta}}^\top A(\boldsymbol{r}+t\boldsymbol{\theta})\Pi_{\boldsymbol{\theta}} dt. && (3.11)
\end{aligned}$$

This is now in the classical TRT format as in Definition 1.4.7.

### 3.6.2 Physical setting of the transverse ray transform

Section 1.4.5 revealed two analytical properties of the TRT from Theorem 1.4.9:

- $\mathrm{TRT}[\vec{U}] = \mathrm{TRT}\left[\vec{U} + [\nabla\varphi]_\times\right]$ for all $\varphi$ such that $\varphi|_{\partial\Omega} = 0$.
- If $\mathrm{TRT}[\vec{U}](\boldsymbol{\theta},\cdot) = \mathrm{TRT}[\vec{V}](\boldsymbol{\theta},\cdot)$ for all $\boldsymbol{\theta}$ on three well-chosen tilt series then $\mathrm{Sym}(\vec{U}) = \mathrm{Sym}(\vec{V})$.

The first null-space will affect all applications but the scan geometry required for exact identification of the symmetric component cannot be achieved in practice. In application of Lemma 3.6.1, signal of the TRT can only be computed if two non-colinear spots are visible in the diffraction pattern. This places a strict constraint on the choice of $\boldsymbol{\theta}$ and, in particular, the set of physically feasible beam orientations is a discrete set. Because of this, datasets in this application will always be in a limited data scenario.

More subtly, the scaling constant switching from average to integral was ignored in (3.11), however, practically this represents a necessary re-scaling by the thickness of the specimen to go from the raw data (average deformation) to the linear model (integral of deformation). This also appears in the analysis as a violation of the small strain assumption, Assumption 6, because the vacuum outside of the crystal has a deformation of order one. This scaling allows us to account for the violation by incorporating outside knowledge of the specimen. Experimentally, a high-angle annular dark-field (HAADF) STEM image can be recorded at each specimen orientation to record the object thickness.

## 3.7 Computational validation

We perform a computational analysis to validate the approximation of the forward model (3.8) used to relate electron diffraction to the TRT and to confirm that the TRT inverse problem, which is under-determined and has a non-trivial null-space, can be solved accurately using a realistic amount of data. These tasks are separated for reasons of efficiency by first rigorously testing the forward model to provide a worse case estimate for the error level and then simulate error at and above this level for numerical tomographic reconstructions.

### 3.7.1 Forward model validation

Physical validation requires quantifying the level of error in (3.8) knowing the exact deformation and comparing against the computed deformation determined based on measuring the centres of the Bragg diffraction disks. This will be assessed in three ways. First we use the MULTEM (Lobato and Van Dyck, 2015) package as a dynamical simulator, in particular modelling multiple scattering effects. We suggest this provides a worst-case analysis as the simulation model is more accurate than the analytical model used, and also we use thick crystals where the extra complexity should be most apparent. Second we compare against a kinematical simulation, the analytical model of this study, and observe that the accuracy is equivalent to that of the dynamical simulation. The previous two comparisons use crystals formed as in (3.3) which ignores strain on the boundaries of each block. The final comparison uses crystals with dislocations where the deformation is continuously defined to verify robustness to the specific definition of strain.

### Piece-wise affine deformation phantom

Phantom crystals are defined with piecewise linear deformations as in (3.3) but using the rigid-ion model of deformation, i.e. where the array of atoms is deformed but atoms remain spherical and the same size. To do this, we create a three parameter collection of phantoms.

**Definition 3.7.1.** *Given an initial crystal structure, randomly sampled phantoms are built up atomistically using three parameters:*

1. *$L \in \mathbb{N}$ is the number of layers. Phantoms consist of layers stacked orthogonal to the $z$-axis. Each layer is under constant affine transformation and is (approximately) the same thickness.*

2. *$d \in \{1, 2, 3\}$ is the rank of the displacement gradient tensor. $d = 1$ corresponds to a simple isotropic scaling of the initial crystal structure. $d = 2$ also allows for rotation and shearing within the layer. $d = 3$ allows any generic $3 \times 3$ tensor.*

3. *$\sigma \geq 0$ is the average magnitude of atomic displacement. In particular, over each randomly generated displacement gradient tensor, the average of the spectral norm of the perturbation from identity will be equal to $\sigma$. Note that this average is not enforced in each simulated phantom.*

### Continuous deformation phantom

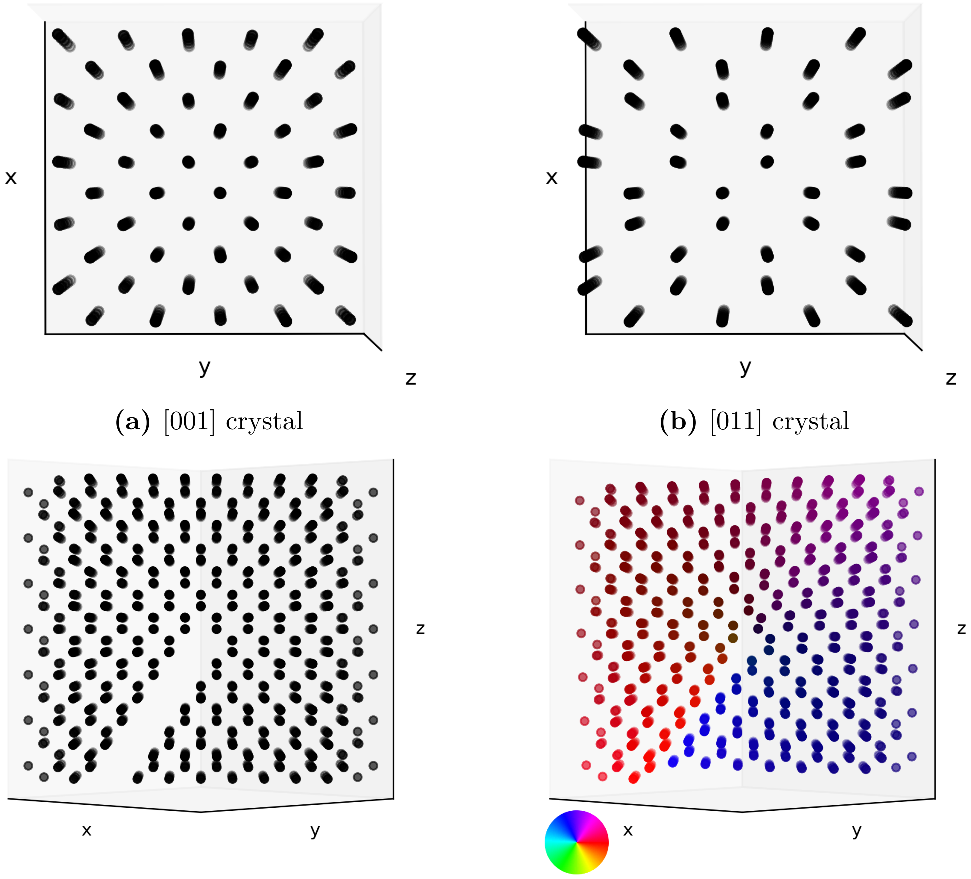
When defined atomistically, the deformation gradient tensor requires a choice of interpolation to be related our discussion in Section 3.3. The piecewise affine deformation phantoms test this in a discrete sense where deformations were defined to be piecewise constant with discontinuities at the interfaces between layers. To test accuracy under continuous deformation we define use the continuum deformation associated with a dislocation with Burger's vector $\boldsymbol{b} \propto (1, 1, 1)^\top$ and line vector $\widehat{\boldsymbol{u}} \propto (1, -1, 0)^\top$ (Hirsch et al., 1967). In the notation of Definition 3.3.1

$$\vec{R}(\boldsymbol{r}) := \frac{4(1-\nu)\theta + \sin(2\theta)}{8\pi(1-\nu)}\boldsymbol{b} + \frac{2(1-2\nu)\log(r) + \cos(2\theta)}{8\pi(1-\nu)}(\widehat{\boldsymbol{u}} \times \boldsymbol{b}) - \boldsymbol{r} \tag{3.12}$$

where $(r, \theta, Z)$ are the cylindrical coordinates of $\boldsymbol{r}$ on the basis $(\widehat{\boldsymbol{u}} \times \boldsymbol{b}, \widehat{\boldsymbol{u}} \times \boldsymbol{b} \times \widehat{\boldsymbol{u}}, \widehat{\boldsymbol{u}})$. A dislocation can be modelled naively by removing all atoms along the half-plane defined by $\theta = \pi$ from a crystal and displacing the atoms according to (3.12). This is visualised in Figure 3.2 and 100 diffraction patterns were simulated from this phantom with beam parallel to the $z$-axis and $\alpha = 2$.

### Electron diffraction simulations

Electron diffraction patters were simulated using a wavelength $\lambda = 0.02$ Å (energy of ca. 300 keV) and an average deformation magnitude of $\sigma = 0.01(= 1\%)$ assuming an incident electron probe



**(a)** [001] crystal

**(b)** [011] crystal

**(c)** Unstrained dislocation

**(d)** Strained dislocation

**Figure 3.2** Parts (a) and (b) show columns of atoms parallel to the beam for two Silicon crystals. To form a dislocation, we remove a half-plane of atoms (c) atoms were displaced according to (3.12) (d). Deformation direction/magnitude is visualised by color/brightness respectively.

function of the form,

$$\Psi_p(x,y) = \frac{J_1(r\sqrt{x^2+y^2})}{r\sqrt{x^2+y^2}}, \qquad \mathcal{F}[\Psi_p](\boldsymbol{k}) = \begin{cases} \frac{1}{r^2} & |\boldsymbol{k}| < r \\ 0 & \text{else} \end{cases}$$

where $J_1$ is a Bessel function of the first kind. This probe function corresponds to the probe produced using a circular disk aperture in ideal probe forming optics and the choice of $r$ corresponds to an aperture of 2 mrad.

Kinematical simulations are computed using (3.2) while dynamical simulations were performed using the multislice approach as implemented in the MULTEM package (Lobato and Van Dyck, 2015). The number of points chosen to simulate precession was chosen sufficiently large such that the errors reported in Table 3.1 had converged, details are given in Section B.5. Precession angles in the range 0° to 2° were used.

### Results: Linearised model accuracy

Computations presented in this section were performed to determine the accuracy with which Bragg disk centres can be measured and related to the average deformation tensor. We also considered how this accuracy may be affected by both the choice of precession angle and the choice of disk centre detection algorithm. The theory suggested in (3.8), that the centre of mass is an accurate measure of average deformation and we compare this with disk-detection by cross-correlation, which is common in strain mapping literature, to evaluate which method is best at linearising the forward model.

Precession angle is an experimental parameter which (3.7) suggests should be chosen as

$$\alpha \approx \cos^{-1}\left(1-\frac{\sigma^2}{2}\right)+\sin^{-1}\left(\frac{\lambda P}{4\pi}\right) \approx 0.01+\frac{0.02\times 5}{4\pi} \approx 0.6^\circ + 0.5^\circ = 1.1^\circ.$$

It is more common in practice to use $\alpha < 1^\circ$, however in this case we wish to quantify any expected advantage of using larger values.

The following simulations were performed:

- 72 MULTEM simulated diffraction patterns with one random phantom for each combination of $\alpha \in \{0^\circ, 0.5^\circ, 1^\circ, 2^\circ\}$, $L \in \{1, 3, 15\}$, $d \in \{1, 2, 3\}$ and crystal orientations [001] and [011] (see Figure 3.2). The full phantom objects were 1000 Å thick.

- 2160 kinematical simulated diffraction patterns with 30 random phantoms for each combination of $\alpha \in \{0^\circ, 0.5^\circ, 1^\circ, 2^\circ\}$, $L \in \{1, 3, 15\}$, $d \in \{1, 2, 3\}$ and crystal directions [001] and [011] parallel to the optic axis ($z$-axis). The full crystals were 250 Å thick.

- 540 high-energy kinematical simulated diffraction patterns without precession and with 30 random phantoms for each combination of $L \in \{1, 3, 15\}$, $d \in \{1, 2, 3\}$ and crystal directions [001] and [011] parallel to the optic axis ($z$-axis). The full crystals were 250 Å thick.

- 100 kinematical simulated diffraction patterns of the dislocation phantom with $\alpha = 2^\circ$ at different beam locations. The full crystal was 250 Å thick.

In particular, we used a disk-detection method involving patch-wise (least squares) registration between each Bragg disk in the strained diffraction pattern and the corresponding Bragg disk in an unstrained diffraction pattern, the exact form of this is in (3.13). A sketch of this pipeline is provided in Figure 3.3.

| Method | dynamical simulation | | | | high-energy |
|---|---|---|---|---|---|
| | $\alpha = 0°$ | $\alpha = 0.5°$ | $\alpha = 1°$ | $\alpha = 2°$ | |
| Centre of Mass | 1.95 | 0.40 | 0.20 | 0.08 | 0.03 |
| Registered | 0.46 | 0.10 | 0.05 | 0.04 | 0.04 |

| Method | kinematical simulation | | | | dislocation phantom |
|---|---|---|---|---|---|
| | $\alpha = 0°$ | $\alpha = 0.5°$ | $\alpha = 1°$ | $\alpha = 2°$ | |
| Centre of Mass | 0.44 | 0.27 | 0.08 | 0.04 | 0.12 |
| Registered | 0.11 | 0.06 | 0.05 | 0.04 | 0.10 |

**Table 3.1** All values given are mean relative Euclidean error (3.14). The dislocation phantom is described in Section 3.7.1, the remainder use layered phantoms. The dynamical simulation is computed by MULTEM, high-energy/kinematical from (3.4)/(3.2).

For each precessed diffraction pattern we compute centres for each spot in the inner-most ring on the pattern. Predicted and computed centres are only ever compared like-for-like relative to centres computed with the same algorithm using an undeformed sample as reference.

In the notation of (3.8), we define the true centres of each spot as

$$\boldsymbol{c}_{true} = \mathop{\mathbb{E}}_{j} \Gamma^\top A_j \boldsymbol{p}_i$$

in the case of discrete piecewise affine deformation, for the continuous deformation map $\vec{R}$

$$\boldsymbol{c}_{true} = \Gamma^\top \boldsymbol{p}_i + \fint_{|x-p_{i,x}|\leq 15\,\text{Å}} \fint_{|y-p_{i,y}|\leq 15\,\text{Å}} \fint_0^T \Gamma^\top \nabla\vec{R}(x,y,z)\boldsymbol{p}_i dz dy dx$$

computes an average deformation where $T$ is the known thickness of the phantom. We then compute the predicted centers

$$\boldsymbol{c}_{com} = \frac{\int_{|\boldsymbol{k}-\Gamma^\top\boldsymbol{p}_i|<\bar{r}} \boldsymbol{k} D_\alpha(\boldsymbol{k}) d\boldsymbol{k}}{\int_{|\boldsymbol{k}-\Gamma^\top\boldsymbol{p}_i|<\bar{r}} D_\alpha(\boldsymbol{k}) d\boldsymbol{k}}, \qquad \boldsymbol{c}_{reg} = \operatorname*{argmin}_{\boldsymbol{p}} \int_{|\boldsymbol{k}-\Gamma^\top\boldsymbol{p}_i|<\bar{r}} |D_\alpha(\boldsymbol{k}) - D_\alpha^0(\boldsymbol{k}+\boldsymbol{p}-\Gamma^\top\boldsymbol{p}_i)|^2 d\boldsymbol{k} \tag{3.13}$$

where $\Gamma^\top\boldsymbol{p}_i$ is computed from an undeformed diffraction pattern, $D_\alpha^0$. We fit a zero-mean Gaussian to the TRT modelling errors and so the important error measure is the Euclidean distance between detected and expected centres, i.e. the error variance. We report the values of

$$\text{error } = 100 \cdot \frac{|\boldsymbol{c}_{true} - \boldsymbol{c}|}{|\boldsymbol{c}_{true}|} \tag{3.14}$$

which is scaled to percentage error. This is convenient because with $\sigma = 1\%$, a naive detection algorithm of $\boldsymbol{c} = \Gamma^\top\boldsymbol{p}_i$ (i.e. zero strain) corresponds to an average error of 1.

Table 3.1 summarises the results of this comparison. The key observations are:

- An average error of one pixel width would correspond to an error of 0.7%, all results below this are super-resolved.

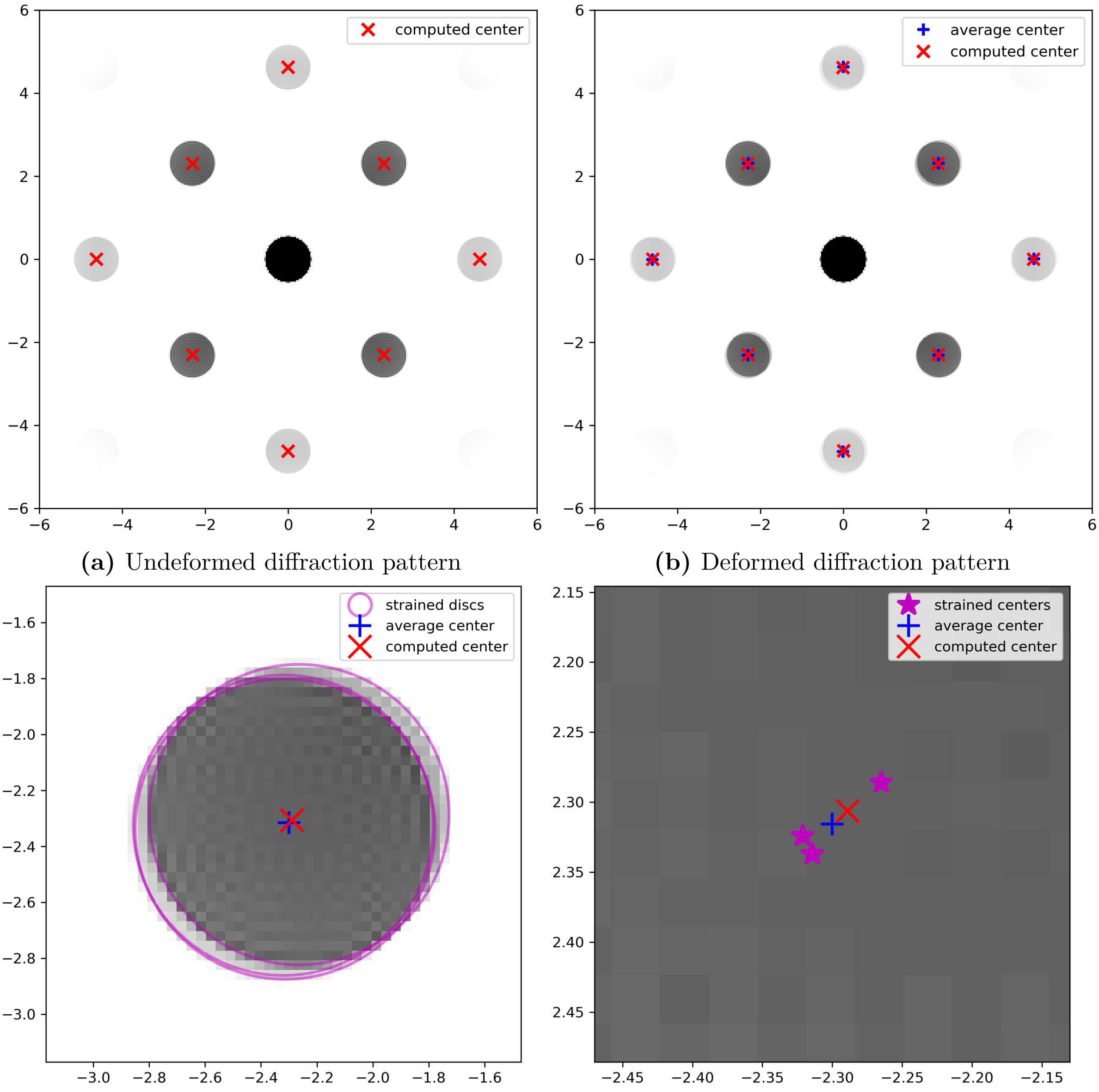
- Increasing the precession angle in this range reduces the errors for all models and centre detection methods.

- Comparing centre detection algorithms, both have comparable maximum accuracy yet the registration method appears much more robust to changes in the simulation mode and phantom. It also converges much faster with respect to precession angle with little gain between 1° and 2°.

- Errors for the continuously deformed phantom are noticeably worse than with the piece-wise constant phantoms. An example disc is shown in Figure B.2 and we see it is qualitatively very different from those shown previously in Figure 3.2. The more smoothly varying deformation causes an elliptical blurring of the disc which may make it harder to consistently detect the centre.

- The high-energy model achieved optimal accuracy without precession. This suggests the dominant benefit of precession in these examples is smoothing the non-linearities of the model rather than accounting for rotation out-of-plane.

In a worst case, with 2° of precession, we observe errors between peak-finding and the TRT model approximation of 0.12%, corresponding to a signal-to-noise ratio (SNR) over 8. On the other hand, the registration peak finding algorithm consistently achieves an average accuracy of 0.04% corresponding to SNR= 25.



**(a)** Undeformed diffraction pattern

**(b)** Deformed diffraction pattern

**(c)** Superimposed deformed discs

**(d)** Distribution of deformed centers

**Figure 3.3** Evaluation pipeline: we compute centres on undeformed (a) and deformed crystals (b). The TRT uses (a) to predict an average center (blue cross) at the center of mass of deformed centers (purple stars). (c) and (d) give sub-plots of (b) to visualise the effect of deformation.

### 3.7.2 Tomographic reconstruction validation

Section 3.6 considers the analytical properties of the continuous inverse problem we wish to solve but in practice we have corrupted and limited data. Here, we perform a reconstruction from a realistic dataset and analyse the accuracy both quantitatively and qualitatively. For this purpose we generate a phantom, $\vec{\mathrm{E}}^\dagger\colon [-1,1]^3 \to \mathbb{R}^{3\times 3}$ and simulate data using the model

$$\vec{\eta}(\boldsymbol{\theta},\boldsymbol{r}) = \mathrm{TRT}[\vec{\mathrm{E}}^\dagger](\boldsymbol{\theta},\boldsymbol{r}) + \Pi_{\boldsymbol{\theta}}\vec{\nu}(\boldsymbol{\theta},\boldsymbol{r})\Pi_{\boldsymbol{\theta}}$$

where $\vec{\nu}$ is a 0-mean isotropic white noise tensor field.

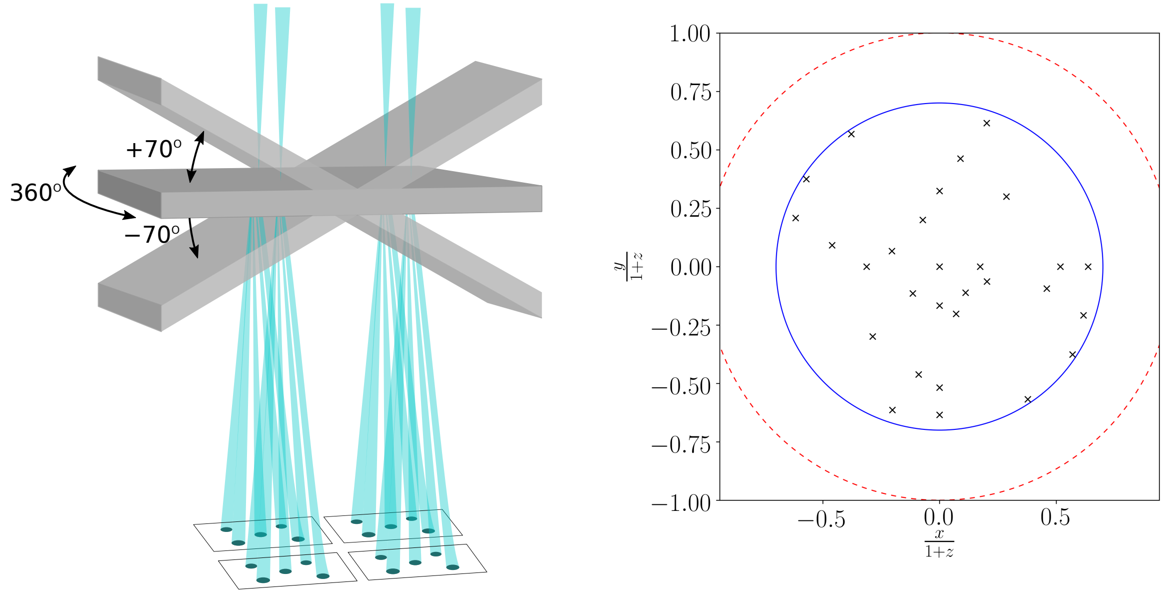
The choice of noise level and phantom were chosen to be slightly more challenging than suggested by the results of Section 3.7.1. In particular, a phantom is chosen with $L = 3$ layers, $d = 3$ dimensional deformation, and an average deformation magnitude of $\sigma = 2\%$. We perform reconstructions with two levels of noise. Firstly at 0.1%, in line with Table 3.1, and then at 1% (SNR $= 2$) to account for any noise and physical modelling errors not considered previously. The final experimental choice is to specify a scan geometry which respects practical time and hardware constraints. 42 tilt directions, shown in Figure 3.4, were selected and for each tilt we scan over a $50 \times 50$ grid of beam positions. Directions are chosen down zone axes to guarantee at least two non-colinear Bragg discs in the diffraction patterns, as required by Lemma 3.6.1. For any rotation $R \in \mathbb{R}^{3\times 3}$ of the sample, the corresponding direction is the third column of $R$, $(x, y, z)^\top$. This direction is then plotted with a stereographic projection at point $\left(\frac{x}{1-z}, \frac{y}{1-z}\right)$. The plot in Figure 3.4.b assumes that the crystal is initially perfectly aligned with the $z$-axis. If this is not true then an offset must be computed and passed on the computation of Euler angles for the specimen holder.

For a reconstruction method we choose to perform a standard total variation reconstruction (Goris et al., 2012; Leary et al., 2013; Collins et al., 2017):

$$\vec{\mathrm{E}}^* = \operatorname*{argmin}_{\vec{\mathrm{E}}} \frac{1}{2}\left\|\mathrm{TRT}[\vec{\mathrm{E}}] - \vec{\eta}\right\|_2^2 + \mu\int_{[-1,1]^3} |\nabla\vec{\mathrm{E}}(\boldsymbol{r})|_{Frobenius} d\boldsymbol{r}$$

where $\mu$ is a manual tuning parameter. With perfect data there is little need for regularisation ($\mu = 0$) however, in this example the total variation functional is compensating for:

- Measurement noise, representing modelling errors at either at 0.1% or 1% magnitude
- Limited angular range, within 70° of the initial orientation
- Limited projections, 42 projections with a $50 \times 50$ grid of beam positions
- Analytical null space, part of the skew component of the tensor field is unobserved by the TRT (i.e. $\mathrm{TRT}[\vec{\mathrm{E}}] = \mathrm{TRT}[\vec{\mathrm{E}} + [\nabla\varphi]_\times]$ for all $\varphi \in C_0^1$) as discussed in Section 3.6.



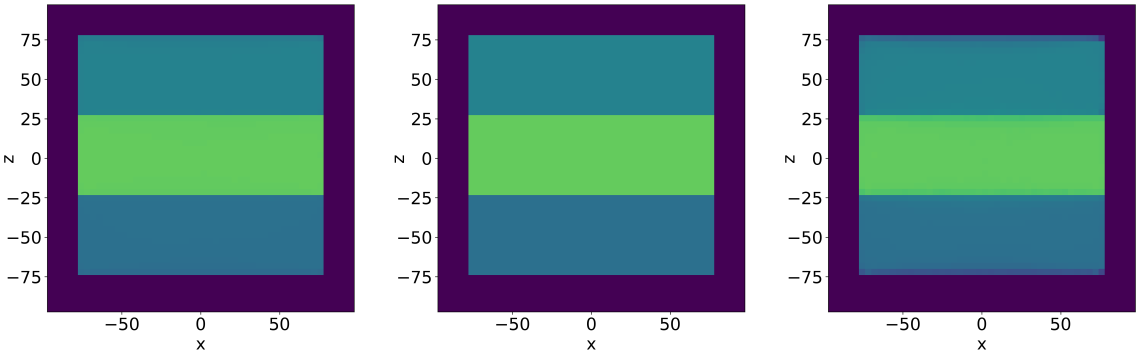
**(a)** Tilt-rotate SPED **(b)** [001] pole figure

**Figure 3.4** Visualisation of acquisition geometry. (a) Tilt-rotate specimen holders allow two degrees of rotation in the form of Euler angles. The 'rotate' angle has full range and the 'tilt' is limited to 70°. (b) Stereographic projection of chosen tilt directions. Crosses indicate 42 zone axes within the limited tilt range of 70° indicated by the solid blue line. The full 90° tilt range is indicated by the dashed red line.

While each of these factors has a different physical or analytical origin, numerically they are all incorporated into a choice of $\mu > 0$. In both reconstructions the parameter was coarsely tuned to $\mu = \frac{10^{-4}}{2}$.

Many other choices of variational methods exist, for instance those compared by Leary et al. (2013), which each account for noise and 'fill in' missing data in their own characteristic fashion. total variation is commonly chosen because it promotes sparse jumps in the reconstruction (Leary et al., 2013; Ehrhardt et al., 2015). This specifically reflects the structure of the phantom in this Section but is also often accurate for other physical samples.

Figure 3.5 visually compares reconstructions from low/high noise data against the original phantom. We see that the general symmetry of the phantom is preserved, reconstructions are uniform in the $x$- and $y$-axes and partition into clear slices along the $z$-axis. The errors in the reconstruction from low noise are imperceptible but at higher noise levels we see the layer interfaces have a slight blur. Figure 3.6 allows us to quantify these errors more precisely. The cross section at low noise shows that the structure of the deformation is well recovered, however, in the flat regions a small consistent error is made in the deformation tensor. Errors at high noise have the same structure but larger magnitude, approximately a factor of 6.5 larger error for a factor of 10 larger noise. The interfaces are still well identified however the jump is more visibly blurred. In all cases, approximately 3 pixels away from a discontinuity errors improve rapidly, by up to a factor of 10. More detailed error comparisons are given in Figure B.4 show that the 99th percentile is a representative reference for the general structure of errors.



**(a)** Low noise reconstruction **(b)** Ground truth **(c)** High noise reconstruction

**Figure 3.5** 2D renders of ground truth and reconstructions. The phantom consists of three deformed layers arranged along the $z$-axis. Deformation tensors are max-projected from 5D to 3D then averaged down the $y$-axis into 2D.

### Interpretation of errors

The three main sources of error are from noise (or modelling error), acquisition geometry, and the null space in the skew component. Figure 3.6 allows us to compare the impact of each of these factors.

Errors from the noise should distribute uniformly over the reconstruction. This error should be constant in uniformly strained regions (near $z = 0$ and $z = \pm 50$) and is visibly much lower than the error at interfaces. Comparing between low and high noise, we also see that errors scale approximately linearly with the noise level.

Tomography with a limited angular range, called limited angle tomography, is common in electron microscopy (Quinto, 1993; Leary et al., 2013; Tovey et al., 2019). From this literature, we know that all changes in the deformation in directions near-orthogonal (at angles greater than 70°) to the $z$-axis are all missing from the observed data. Uncertainty over where jumps occur lead to blurred edges which are seen as the spreading of errors in the $z$ direction. The jump from crystal to vacuum provides a worst-case scenario and, as previously commented, this blurring is approximately 3-4 pixels in radius. This radius is consistent between different noise levels and the full/symmetric strain components.

Comparing the second and third columns of Figure 3.6, it is clear that the error on the symmetric components is no smaller than the full error. This indicates that the contribution of error from the null space of the skew component is negligible.

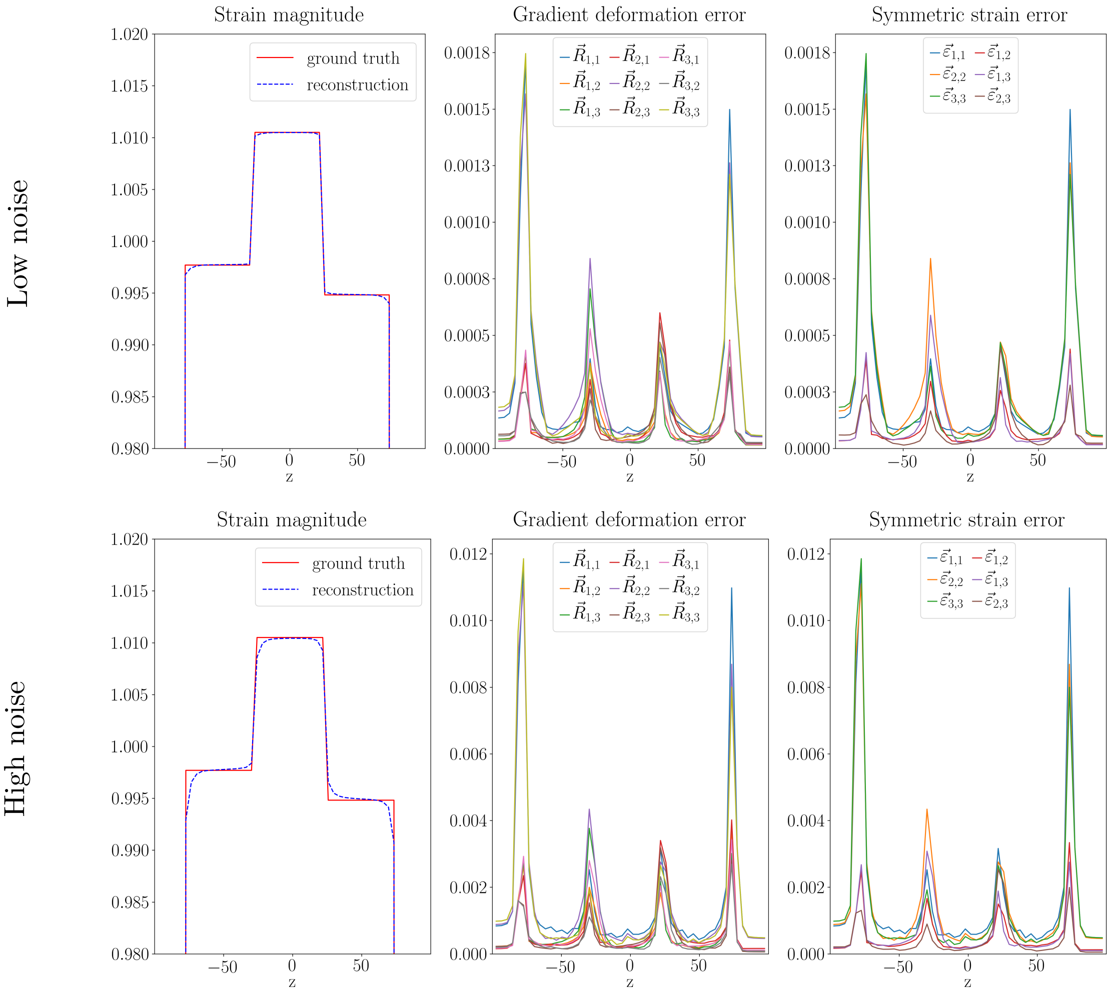


**Figure 3.6** 1D projections of reconstruction volumes and error distributions. In the volume, tensors are projected to scalars by selecting the maximum. For errors, the 99th percentile of each nine components of the gradient deformation tensor (or six components of symmetric strain) are plotted.

Returning to the question of symmetric strain reconstruction, there are the two conventions for the definition of strain, either

$$\overrightarrow{\varepsilon} = \frac{1}{2}(\overrightarrow{R} + \overrightarrow{R}^\top) \qquad \text{or} \qquad \overrightarrow{\varepsilon} = \sqrt{\overrightarrow{R}^\top \overrightarrow{R}}.$$

In both cases they are easily computed from the full displacement gradient tensor $\overrightarrow{\mathrm{E}}^*$. The analysis of Section 3.6 nicely aligns with the first definition, which is the default in this work, but the error analysis above shows that this does not bias the quality of the reconstruction. Both definitions of symmetric strain are reconstructed with equivalent levels of accuracy in this example. The total accuracy is dictated much more by the acquisition geometry than noise (representing modelling error) or missing data in the skew component.

## 3.8 Conclusions and outlook

In this work we have proposed a tomographic model for the 3D strain mapping problem, analysed and extended the known analytical properties of the resulting inverse problem, and provided numerical results for both modelling and inversion steps. Table 3.1 shows that our forward model is accurate up to an SNR of 20 with respect to the TRT model. Beam precession is key to this accuracy and disk registration based methods are more stable to this uncertainty relative to centres of mass. We note that the 0.08% found here agrees exactly with a comparable quantity of Mahr et al. (2015) despite very different deformations considered in each study. In Figure 3.6 we have shown that reconstructions can be performed with realistic experimental parameters and achieve accurate results, even with errors much larger than predicted. Because of this, it is our belief that there is no benefit to quantifying errors beyond this SNR of 20 within the scope of this work, namely the validity of the dynamical model tested in Section 3.7. Realistic reconstructions can be recovered well at this level of error, although this does not take into account many factors such as detector performance, electron optical aberrations, and inelastic scattering. These factors could potentially be the dominant causes of reconstruction error in practice and should be minimised experimentally and assessed further.

Our proposed framework requires diffraction patterns to be recorded near zone axes where diffraction patterns have straight lines of spots. It would be much faster experimentally to acquire many diffraction patterns with single arced lines of spots, called ‘off-axis’ diffraction patterns, however this would take us away from the TRT model. On the theoretical side, there have been recent advances in histogram tomography which could assist the well-posedness of strain tomography (Lionheart, 2019). In particular, in this study we compute the centre of mass (first moment) of each spot in a diffraction pattern and the resulting inverse problem is the TRT which always has a large null-space. If we also extracted second order moments from each spot then the model is no longer the TRT and the extra data may remove the issue of non-unique solutions. Finally, there is an interesting conflict in the desired scan geometry due to the physical model and the TRT. As commented in Section 3.6.2, theoretical results of the TRT rely on three orthogonal tilt series whereas we only have access to data at a discrete set of orientations which, dependent on the crystal structure, arise (approximately) uniformly over the sphere. It would be interesting to unify these two pressures and analyse characteristics of the tomography problem in such a constrained geometry.

# Chapter 4

# FISTA with Adaptive Discretisation

In many variational approaches to inverse problems, optimisation problems are often written in a continuous setting but solved with a discrete optimisation algorithm. Consider what happens when $\sqrt{e^2}$ is computed in your favourite programming language. We have asked a 'continuous' question and have been given a 'discrete' answer with very well controlled error bounds.

The parallel to this for optimisation is that we still do not have infinite computing memory or processing power, but we still want to compute minimisers to a known accuracy with efficient implementations. It is typically much harder to control discretisation errors in optimisation problems, for instance the $\Gamma$-convergence of discrete total variation (Bartels, 2012, 2015) or the approximation errors for wavelets/curvelets/sheerlets (Mallat, 1999; Candès and Donoho, 2004; Guo and Labate, 2007). These are typically quite weak guarantees valid for asymptotic resolution/time and are hard to quantify in a specific example. In this work we propose an algorithm with three aims:

- All computations are discrete but the asymptotic reconstruction is exact, even for infinite dimensional reconstructions in a Banach space, with a guaranteed rate.
- Distance to the infinite dimensional minimiser can be quantified.
- Discretisation is adaptively optimised for the particular minimiser.

We choose to modify the FISTA algorithm because it is quite general and performs very well in discrete optimisation problems. The strategy will be to follow the standard FISTA algorithm as closely as possible. The only difference is that at each iteration, computations will only be performed on a finite dimensional subspace.

One key issue is that FISTA is intrinsically designed to converge only in $L^2$. In finite dimensions there is no issue, all norms are equal and so convergence rates are equal for all problems up to scaling constants. On the other hand, in infinite dimensions there exist minimisers which are not in $L^2$. In this case, we find that the rate will always be slower than for standard FISTA but still at a guaranteed rate.

We perform two numerical experiments where the minimiser is in $L^1 \setminus L^2$ which demonstrates the reduced rate. We also show an example where the minimiser is in $\ell^2(\mathbb{R}^m)$ for some $m$ which is not known a priori and observe that the refining FISTA algorithm achieves linear convergence.

## 4.1 Introduction

The standard setting of the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) is that there exists a Hilbert space $\mathbb{H}$ over which we wish to minimise the function

$$\min_{u\in\mathbb{H}} \mathrm{E}(u) \qquad \text{such that} \qquad \mathrm{E}(u) \coloneqq \mathrm{f}(u) + \mathrm{g}(u), \tag{4.1}$$

where $\mathrm{f}\colon \mathbb{H} \to \mathbb{R}$ is a convex differentiable function with $L$-Lipschitz gradient and $\mathrm{g}\colon \mathbb{H} \to \overline{\mathbb{R}}$ is a convex function. FISTA is a practically fast algorithm which, for many choices of E, generates a sequence of iterates $u_n \in \mathbb{H}$ such that $\mathrm{E}(u_n)$ converges at the optimal rate of $O(n^{-2})$ (Beck and Teboulle, 2009; Chambolle and Dossal, 2015).

The canonical example for this work will be the Lasso energy,

$$\mathrm{E}(u) = \tfrac{1}{2}\left\|\mathcal{A}u - \boldsymbol{\eta}\right\|_2^2 + \mu\left\|u\right\|_1$$

where $\boldsymbol{\eta} \in \mathbb{R}^m$ is some observed data, $\mathcal{A}\colon \mathcal{M}(\mathbb{R}^d) \to \mathbb{R}^m$ is the forward map and $\mu > 0$ is a chosen scalar. It is known that there exist minimisers of E of the form:

$$u^* = \sum_{i=1}^{m} \alpha_i \delta_{\boldsymbol{x}_i},$$

for some $\alpha_i \in \mathbb{R}$ and where $\delta_{\boldsymbol{x}_i}$ is the Dirac delta function centred at $\boldsymbol{x}_i \in \mathbb{R}^d$ (Unser et al., 2016; Boyer et al., 2019). The challenge with this minimiser is that $u^* \in L^1 \setminus L^2$, so exact FISTA cannot be applied to this problem. On the other hand, $u^*$ has a very nice structure which we expect can be easily and efficiently represented by, for instance, a basis of finite elements.

In this work we propose a modification of FISTA which addresses both of these points. During each iteration we restrict computation to a subspace, i.e. $u_n \in \mathbb{U}^n \subset \mathbb{H}$. For infinite dimensional optimisation, this allows us to reconstruct $u^*$ to an arbitrary precision in finite time. Alternatively, for finite dimensional optimisation this potentially allows for a more computationally efficient variant of the classical FISTA algorithm.

Inexact optimisation is a well-established field which can be seen to encompass methods such as coordinate descent (Wright, 2015), stochastic gradient (Spall, 2005; Bottou et al., 2018), or indeed approximate FISTA-like algorithms (Jiang et al., 2012; Villa et al., 2013). The result of Theorem 4.4.8 is very similar to (Schmidt et al., 2011, Proposition 2) and (Aujol and Dossal, 2015, Proposition 3.3). The key novelty is the concept of convergence outside of the Hilbert space $\mathbb{H}$. Additionally assuming that errors come from quantified subspace approximations, we can also provide explicit rates.

### 4.1.1 Outline

This chapter is organised as follows. Section 4.2 defines notation and the generic form of our proposed refining FISTA algorithm, Algorithm 4.1. The main theoretical contribution of this work is the convergence analysis of Algorithm 4.1 which is split into two parts: first we outline the proof structure in Section 4.3, then we state the specific results in the case of FISTA in Section 4.4. The main results are Theorem 4.4.8 and Lemma 4.4.11 which extend the convergence of FISTA to infinite dimensional Banach spaces with uniform/adaptively chosen subspaces $\mathbb{U}^n$ respectively.

Section 4.5 presents some general results for the application of Algorithm 4.1 and Section 4.6 gives a much more detailed discussion of adaptive refinement for Lasso minimisation. In particular, we describe how to choose efficient refining discretisations to approximate $u^*$, estimate the convergence of E, and identify the support of $u^*$. The numerical results in Section 4.7 demonstrate these techniques in four different models demonstrating the sharpness of our analysis and the computational efficiency of adaptive discretisations.

## 4.2 Definitions and notation

We consider optimisation of (4.1) over two spaces, the Banach space $(\mathbb{U}, |||\cdot|||)$ and Hilbert space $(\mathbb{H}, \langle\cdot,\ \cdot\rangle, \|\cdot\|)$, such that

$$\exists u^* \in \mathbb{U} \qquad \mathrm{s.\,t.} \qquad \mathrm{E}(u^*) = \min_{u\in\mathbb{U}} \mathrm{E}(u) = \min_{u\in\mathbb{H}} \mathrm{E}(u).$$

We further define

$$\mathrm{E}_0 \colon u \mapsto \mathrm{E}(u) - \mathrm{E}(u^*).$$

Note that $\mathrm{E}_0$ is uniquely defined, even if $u^*$ may not be.

We propose a refining FISTA algorithm in Algorithm 4.1 for a choice of refining subspaces $\mathbb{U}^n \subset \mathbb{U}^{n+1} \subset \mathbb{U}\cap\mathbb{H}$ for $n = 0, 1, \ldots$. The only difference is that on iteration $n$, all computations are performed in the subspace $\mathbb{U}^n$. Indeed, if $\mathbb{U}^n = \mathbb{U} = \mathbb{H}$ then this is just the standard FISTA algorithm.

Without loss of generality we will assume $L = 1$, i.e. $\nabla \mathrm{f}$ is 1-Lipschitz. To get the general statement of any of the results which follow, replace E with $\frac{\mathrm{E}}{L}$. Implicitly, the only impact of $L$ is scaling the convergence rates which are only given up to a constant factor anyway. A more important distinction is to say that the properties of f and g are stated with respect to the Hilbert norm. In particular,

$$\|\nabla \mathrm{f}(u) - \nabla \mathrm{f}(v)\| \leq \|u - v\|$$

for all $u, v \in \mathbb{H}$ and g is called 'simple' if

$$\operatorname*{argmin}_{u\in\widetilde{\mathbb{U}}} \tfrac{1}{2}\left\|u-v\right\|^2 + \mathrm{g}(u)$$

is exactly computable for all $v \in \mathbb{H}$ and all $\widetilde{\mathbb{U}} \in \{\mathbb{U}^n\}_{n=0}^\infty$.

Finally we introduce the concept of an orthogonal projection in the setting of this work. For a subspace $\mathbb{U}^n \subset \mathbb{U} \cap \mathbb{H}$ we define the orthogonal projection $\Pi_n\colon (\mathbb{U}^n)^* \to \mathbb{U}^n$ to be any extension of the function such that

$$\langle \Pi_n u,\ v\rangle = \langle u,\ v\rangle \text{ for all } u \in (\mathbb{U}^n)^*,\ v \in \mathbb{U}^n.$$

This definition is non-standard as we view $(\mathbb{U}^n)^*$ embedded in $\mathbb{U}$. Note that $\mathbb{U}^n \subset \mathbb{H}$ implies that $(\mathbb{U}^n)^* \supset \mathbb{H}$, therefore $\Pi_n$ is an extension of the classical orthogonal projection. It is convenient to allow an extension here because $\mathbb{U}$ may be bigger than $\mathbb{H}$, $(\mathbb{U}^n)^*$ is the largest space such that $\Pi_n$ is well-defined. Beyond this, the Hahn-Banach theorem may allow the domain of $\Pi_n$ to be extended further, but the definition is no longer unique. For example, in 1D the value of $\left\langle \delta_0,\ \frac{1}{\sqrt[4]{x}}\right\rangle$ is not well-defined, but the value of $\left\langle \delta_0,\ \mathbb{1}_{[0,1]}\right\rangle$ can be chosen to be consistent with some sequence of molifiers. To account for this possibility, we shall state that the domain of $\Pi_n$ is $\overline{(\mathbb{U}^n)^*}$, with the closure computed in an appropriate topology dictated by $|\!|\!|\cdot|\!|\!|$.

Constant factors will generally not be tracked during the proofs in this chapter. For sequences $(a_n)_{n=1}^\infty$,$(b_n)_{n=1}^\infty$ we will frequently use the notation:

$$\begin{aligned} a_n \lesssim b_n \qquad &\iff \qquad \exists C, N > 0 \text{ s. t. } a_n \leq C b_n \text{ for all } n > N,\\ a_n \simeq b_n \qquad &\iff \qquad a_n \lesssim b_n \lesssim a_n. \end{aligned}$$

**Algorithm 4.1** Refining sub-space FISTA

```
Choose $(\mathbb{U}^n)_{n\in\mathbb{N}}$, $u_0 \in \mathbb{U}^0$ and some FISTA stepsize choice $(t_n)$
$v_0 \leftarrow u_0, n \leftarrow 0$
repeat
    $\overline{u}_n \leftarrow (1-\frac{1}{t_n})u_n + \frac{1}{t_n}v_n$
    $u_{n+1} \leftarrow \operatorname*{argmin}_{u\in\mathbb{U}^{n+1}} \frac{1}{2}\left\|u-\overline{u}_n + \nabla \mathrm{f}(\overline{u}_n)\right\|^2 + \mathrm{g}(u)$
    $v_{n+1} \leftarrow (1-t_n)u_n + t_n u_{n+1}$
    $n \leftarrow n+1$
until converged
```

## 4.3 General proof recipe

In this work we focus on the FISTA algorithm, however, the key ingredients of the proof do not rely on the particular structure of FISTA. In this section we will sketch the general 'recipe' of the convergence proof for adaptive schemes in a Banach space setting.

During this section, we will refer to the structure of FISTA as motivation. In particular, we recall the classical FISTA convergence guarantee given in (Chambolle and Dossal, 2015, Theorem 2):

$$t_N^2\,\mathrm{E}_0(u_N) + \sum_{n=1}^{N-1} \rho_n\,\mathrm{E}_0(u_n) + \tfrac{1}{2}\left\|v_N - u^*\right\|^2 \leq \tfrac{1}{2}\left\|u_0 - u^*\right\|^2 \tag{4.2}$$

for some $t_N \simeq N$ and $\rho_n \geq 0$.

#### Step 1: Quantifying the scaling properties

The first step is to quantify how E and $\|\cdot\|$ behave as the discretisation refines, or resolution increases. In Algorithm 4.1 we are given the subspaces $\mathbb{U}^n$ which we partition into a sequence of milestones. In particular, we assume there exists $n_k \in \mathbb{N}$ and constants $a_U, a_{\mathrm{E}} \geq 1$ such that:

$$n_0 < n_1 < \ldots, \qquad \left\|\operatorname*{argmin}_{u\in\mathbb{U}^{n_k}} \mathrm{E}_0(u)\right\| \lesssim a_U^k, \qquad \text{and} \qquad \min_{u\in\mathbb{U}^{n_k}} \mathrm{E}_0(u) \lesssim a_{\mathrm{E}}^{-k}.$$

The idea is that $\mathbb{U}^{n_k}$ is a discretisation at resolution $h^k$ for some $h < 1$ and therefore the minimum of $\mathrm{E}_0$ decays exponentially while the norm of discrete minimisers potentially grows exponentially. In Section 4.5 we will see that this exponential scaling is very natural.

The value of $a_U$ is dictated by the Banach space $\mathbb{U}$ in relation $\mathbb{H}$. If $u^* \in \mathbb{U}\setminus\mathbb{H}$, as for Lasso, then the right-hand side of (4.2) becomes infinity. All norms are equivalent on finite dimensional subspaces, but the scaling of this relationship is quantified by $a_U > 1$. The value of $a_{\mathrm{E}} > 1$ is an indicator of how easy E is to discretise. If E is very smooth and the choice of discrete basis is very efficient then $a_{\mathrm{E}}$ is large. The trade-off between $a_{\mathrm{E}}$ and $a_U$ dictates the final convergence rate of the algorithm.

#### Step 2: Generalising the convergence bound

The bound in (4.2) is only valid when $\mathbb{U}^n$ is a constant sequence. The first analytical step is to quantify the effect of refinement. Theorem 4.4.4 gives an expression for this for generic choices of $\mathbb{U}^n$ in Algorithm 4.1.

If $\mathbb{U}^n$ is a constant sequence then Theorem 4.4.4 recovers the right-hand side of (4.2) through a large telescoping sum, the exact form of the sum does not matter at this moment. Without further assumptions, the sum does not telescope and the bound grows linearly with $n$. If we re-introduce the sequence $n_k$ from the previous step then we can simplify this inequality. In particular, if

$$\mathbb{U}^{n_k} = \mathbb{U}^{n_k+1} = \ldots = \mathbb{U}^{n_{k+1}-1},$$

then the right-hand side will telescope on the intervals $(n_k, n_{k+1})$ and scale only with $k \simeq \log(n)$. The result of this is presented in Lemma 4.4.5.

The take-home message here is that the introduction of the milestones $n_k$ greatly simplifies the convergence bound expression and allows us to utilise the scaling properties described in Step 1.

### Step 3: Sufficiently fast refinement

In Step 2 we developed a convergence bound, now we wish to show that it is only worse than the classical (4.2) by a constant factor. In particular, it is equivalent to run Algorithm 4.1 for $N$ iterations or the classical FISTA algorithm for $N$ iterations on the fixed subspace $\mathbb{U}^N$. Lemma 4.4.6 shows that this is true so long as $\mathbb{U}^n$ refine sufficiently quickly, i.e. $n_k$ are sufficiently small. In summary, in comparison with (4.2), we show

$$\mathrm{E}_0(u_N) \lesssim \frac{a_U^{2K}}{N^2} = O\left(\frac{\|u_0 - \operatorname{argmin}_{u\in\mathbb{U}^N} \mathrm{E}(u)\|^2}{N^2}\right)$$

for all $N \leq n_K$.

### Step 4: Sufficiently slow refinement

The result of Step 3 is sufficient to prove convergence, but not directly a rate. If the subspaces refine too quickly then this factor of $\|u^*\| = \infty$ will slow the rate of convergence. Refinement should happen sufficiently quickly so that we do not waste time overfitting to the discretisation, but low resolution problems converge faster therefore we should then refine as slowly as possible. Lemma 4.4.7 balances these two factors in an optimal way for the FISTA algorithm. This results in a convergence guarantee of the form

$$\mathrm{E}_0(u_N) \lesssim \frac{N^{2\varepsilon}}{N^2}$$

for all $N \in \mathbb{N}$, some $\varepsilon \in [0, 1)$ depending on $a_U, a_{\mathrm{E}}$. In particular, if $u^* \in \mathbb{H}$ then $\varepsilon = 0$ recovers the classical rate.

### Step 5: Adaptivity

Up to this point we have implicitly focused on the case where $\mathbb{U}^n$ and $n_k$ are chosen a priori. Here we emphasise some of the challenges which are faced when extending results to allow for on-the-fly greedy adaptivity.

Spatial adaptivity is robust, so long as the scaling properties of Step 1 are satisfied. The only other constraint is to ensure that the partial telescoping of Step 2 still holds. In the case of FISTA, Lemma 4.4.5 shows that this only requires the existence of $\widetilde{w}_k \in \mathbb{U}^{n_k}$ such that

$$\widetilde{w}_k \in \mathbb{U}^{n_k} \cap \mathbb{U}^{n_k+1} \cap \ldots \cap \mathbb{U}^{n_{k+1}-1}.$$

Refinement time adaptivity is more challenging for FISTA due to the non-descent property of the algorithm. The idea is that resolution should increase rapidly while E is 'easy' to optimise then default to the rate of Step 4 when it is 'hard'. The result of this is Theorem 4.4.9 which shows

$$\min_{n\leq N} \mathrm{E}_0(u_n) \lesssim \frac{N^{2\varepsilon}}{N^2}$$

for all $N \in \mathbb{N}$, the same $\varepsilon \in [0,1)$ from Step 4. The penalty for accelerating the refinement time is a potential loss of stability in $\mathrm{E}(u_n)$, however, the asymptotic rate is equivalent and this behaviour has not been seen in numerical experiments.

## 4.4 Proof of convergence

In this section we follow the recipe motivated in Section 4.3 to prove convergence of two variants of Algorithm 4.1. Motivated by this argument, we will first formalise the definition of the constants $a_U$ and $a_{\mathrm{E}}$.

**Definition 4.4.1.** *Fix $a_U, a_{\mathrm{E}} \geq 1$ and a sequence of subspaces $\{\widetilde{\mathbb{U}}^k \subset \mathbb{H} \cap \mathbb{U} \text{ s.t. } k \in \mathbb{N}\}$. We say that $\{\widetilde{\mathbb{U}}^k\}$ is a $(a_U, a_{\mathrm{E}})$-*discretisation for E *if*

$$\|\widetilde{w}_k\| \lesssim a_U^k \qquad \textit{and} \qquad \mathrm{E}_0(\widetilde{w}_k) \lesssim a_{\mathrm{E}}^{-k}$$

*for all $k \in \mathbb{N}$ and some choice $\widetilde{w}_k \in \operatorname{argmin}_{u\in\widetilde{\mathbb{U}}^k} \mathrm{E}(u)$.*

In this section we will simply assume that such sequences exist and in Section 4.5 we will give some more general examples. Each of the main theorems and lemmas will be stated with a sketch proof in this section. The details of the proofs are either trivial or very technical and are therefore placed in Section C.1 to preserve the flow of the argument.

### 4.4.1 Computing the convergence bound

For Step 2 of Section 4.3 we look to replicate the classical bound of the form in (4.2) for Algorithm 4.1. The proofs in this step follow the classical arguments of Beck and Teboulle (2009); Chambolle and Dossal (2015) very closely.

### Single iterations

We first wish to understand a single iteration of Algorithm 4.1. This is done through the following two lemmas.

**Lemma 4.4.2** (equivalent to (Chambolle and Dossal, 2015, Lemma 1))**.** *Suppose $\nabla \mathrm{f}$ is 1-Lipschitz, for any $\overline{u} \in \mathbb{U}^{n-1}$ define*

$$u \coloneqq \operatorname*{argmin}_{u\in\mathbb{U}^n} \tfrac{1}{2}\left\|u - \overline{u} + \nabla \mathrm{f}(\overline{u})\right\|^2 + \mathrm{g}(u).$$

*Then, for all $w \in \overline{(\mathbb{U}^n)^*} \supset \mathbb{H}$, we have*

$$\mathrm{E}(u) + \tfrac{1}{2}\left\|u - \Pi_n w\right\|^2 \leq \mathrm{E}(\Pi_n w) + \tfrac{1}{2}\left\|\overline{u} - \Pi_n w\right\|^2$$

*where $\Pi_n \colon \overline{(\mathbb{U}^n)^*} \to \mathbb{U}^n$ is the orthogonal projection.*

This is exactly the result of Chambolle and Dossal (2015) applied to the function $u \mapsto \mathrm{E}(\Pi_n u)$. Applying Lemma 4.4.2 to the iterates from Algorithm 4.1 gives a more explicit inequality.

**Lemma 4.4.3** (analogous to (Chambolle and Dossal, 2015, Theorem 2), (Beck and Teboulle, 2009, Theorem 1))**.** *Let $\mathbb{U}^n \subset \mathbb{H}\cap\mathbb{U}$ and $w_n \in \mathbb{U}^n$ be chosen arbitrarily and $u_n/v_n$ be generated by Algorithm 4.1 for all $n \in \mathbb{N}$. For all $n \in \mathbb{N}$ it holds that*

$$t_n^2(\mathrm{E}(u_n)-\mathrm{E}(w_n))-(t_n^2-t_n)(\mathrm{E}(u_{n-1})-\mathrm{E}(w_n)) \leq \tfrac{1}{2}\left[\left\|v_{n-1}\right\|^2 - \left\|v_n\right\|^2\right]+\langle v_n - v_{n-1},\ w_n\rangle\,. \quad (4.3)$$

The proof of this lemma is a result of the convexity of E for a well chosen $w$ in Lemma 4.4.2.

### Generic convergence bound

Lemma 4.4.3 gives us an understanding of a single iteration of Algorithm 4.1, summing over $n$ then gives our generic convergence bound for any variant of Algorithm 4.1.

**Theorem 4.4.4.** *Fix a sequence of subspaces $\{\mathbb{U}^n \subset \mathbb{U}\cap\mathbb{H} \text{ s.t. } n \in \mathbb{N}\}$, arbitrary $u_0 \in \mathbb{U}^0$, and FISTA stepsize choice $(t_n)_{n\in\mathbb{N}}$. Let $u_n$ and $v_n$ be generated by Algorithm 4.1. Then, for any choice of $w_n \in \mathbb{U}^n$ and $N \in \mathbb{N}$ we have*

$$\begin{aligned} t_N^2\,\mathrm{E}_0(u_N) + \sum_{n=1}^{N-1}\rho_n\,\mathrm{E}_0(u_n) + \frac{\left\|v_N - w_N\right\|^2}{2} \leq{}& \frac{\left\|u_0 - w_0\right\|^2 - \left\|w_0\right\|^2 + \left\|w_N\right\|^2}{2} \\ &+ \sum_{n=1}^{N} t_n\,\mathrm{E}_0(w_n) + \langle v_{n-1},\ w_{n-1} - w_n\rangle\,. \quad (4.4)\end{aligned}$$

This result is the key approximation for showing convergence of FISTA with refining subspaces. In the classical setting, we have $\mathbb{U}^n = \mathbb{U} = \mathbb{H}$, $w_n = u^*$ and the extra terms on the right-hand side collapse to 0.

### Convergence bound with milestones

In standard FISTA, the right-hand side of (4.4) is a constant. The following lemma minimises the growth of the 'constant' as a function of $N$ by partially telescoping the sum on the right-hand side.

**Lemma 4.4.5.** *Let $u_n$, $v_n$ be generated by Algorithm 4.1, $(n_k \in \mathbb{N})_{k=0}^\infty$ be a monotone increasing sequence, and define*

$$\widetilde{\mathbb{U}}^k := \mathbb{U}^{n_k}, \qquad \widetilde{w}_k \in \operatorname*{argmin}_{u \in \widetilde{\mathbb{U}}^k} \mathrm{E}(u).$$

*If*

$$\widetilde{w}_k \in \mathbb{U}^n \qquad \text{for all} \qquad n_k \leq n < n_{k+1},\ k \in \mathbb{N},$$

*then for all $K \in \mathbb{N}$, $n_K \leq N < n_{K+1}$ we have*

$$t_N^2 \, \mathrm{E}_0(u_N) + \sum_{n=1}^{N-1} \rho_n \, \mathrm{E}_0(u_n) + \frac{\|v_N - \widetilde{w}_K\|^2}{2} \leq C + \frac{\|\widetilde{w}_K\|^2}{2} + \frac{(N+1)^2 - n_K^2}{2} \, \mathrm{E}_0(\widetilde{w}_K)$$
$$+ \sum_{k=1}^{K} \frac{n_k^2 - n_{k-1}^2}{2} \, \mathrm{E}_0(\widetilde{w}_{k-1}) + \langle v_{n_k - 1},\ \widetilde{w}_k - \widetilde{w}_{k+1} \rangle$$

*where $C = \frac{\|u_0 - \widetilde{w}_0\|^2 - \|\widetilde{w}_0\|^2}{2}$.*

The introduction of $n_k$ has greatly simplified the expression of Theorem 4.4.4. On the right-hand side, we now only consider $\mathrm{E}_0$ evaluated on the sequence $\widetilde{w}_k$ and there are only $K$ non-zero inner-product terms remaining.

## 4.4.2 Refinement without overfitting

The result of Lemma 4.4.5 is still optimal in the sense that it reduces to (4.2) when $\mathbb{U}^n = \mathbb{U}$, however, now we would like to achieve the equivalent rate (up to a constant factor) including refinement. This can be likened to the idea of overfitting to the discretisation. It is only efficient to optimise the discrete energy while the discrete gap $\mathrm{E}(u_n) - \mathrm{E}(\widetilde{w}_k)$ is comparable to the continuous gap $\mathrm{E}(u_n) - \mathrm{E}(u^*)$.

This is achieved by two assumptions, first we use the structure of Definition 4.4.1 to quantify the properties of the refinement, then we force $K$ to scale with $\log(N)$ to slow the growth of the right hand side. This is summarised in the following lemma.

**Lemma 4.4.6.** *Suppose $\mathbb{U}^n, u_n, v_n$ and $n_k$ satisfy the conditions of Lemma 4.4.5 and $\{\widetilde{\mathbb{U}}^k\}$ forms an $(a_U, a_{\mathrm{E}})$-discretisation for $\mathrm{E}$. If either:*

- *$a_U > 1$ and $n_k^2 \lesssim a_{\mathrm{E}}^k a_U^{2k}$,*
- *or $a_U = 1$, $\sum_{k=1}^\infty n_k^2 a_{\mathrm{E}}^{-k} < \infty$ and $\sum_{k=1}^\infty \|\widetilde{w}_k - \widetilde{w}_{k+1}\| < \infty$,*

*then*

$$\mathrm{E}_0(u_N) \lesssim \frac{a_U^{2K}}{N^2}$$

*for all* $n_K \leq N < n_{K+1}$.

We make two observations of the optimality of Lemma 4.4.6:

- The convergence guarantee for $N$ iterations of classical FISTA in the space $\mathbb{U}^N$ is

  $$\mathrm{E}_0(u_N) \lesssim \frac{\|u_0 - \operatorname{argmin}_{u\in\mathbb{U}^N} \mathrm{E}(u)\|^2}{N^2} + \min_{u\in\mathbb{U}^N} \mathrm{E}(u) \lesssim \frac{a_U^{2K}}{N^2} + a_{\mathrm{E}}^{-K}.$$

  This is equivalent to Lemma 4.4.6 under the conditions on $n_k$.
- If $\mathbb{U}$ is finite dimensional then the $a_U = 1$ is almost trivially satisfied. Norms in finite dimensions are equivalent and any discretisation can be achieved with a finite number of refinements (i.e. the sums over $k$ are finite).

### Convergence rate

At the end of Step 3 we have shown that $\mathrm{E}(u_n)$ converges at a rate depending on $k$ and $n$ so long as $k$ grows sufficiently quickly. On the other hand, as $k$ grows, the rate becomes worse and so we need to also put a lower limit on the growth of $n_k$. The following lemma computes the global convergence rate of $\mathrm{E}(u_n)$ when $k$ grows at the minimum rate which is consistent with Lemma 4.4.6.

As a special case, note that if $a_U = 1$ then Lemma 4.4.6 already gives the optimal $O(N^{-2})$ convergence rate. This is in fact a special case of Aujol and Dossal (2015). If $u^* \in \mathbb{H}$ then it is not possible to refine 'too quickly' and the following lemma is not needed.

**Lemma 4.4.7.** *Suppose $u_n$ and $n_k$ are sequences satisfying*

$$\mathrm{E}_0(u_N) \lesssim \frac{a_U^{2K}}{N^2} \qquad \textit{where} \qquad n_K^2 \gtrsim a_{\mathrm{E}}^K a_U^{2K},$$

*then*

$$\mathrm{E}_0(u_N) \lesssim \frac{1}{N^{2(1-\varepsilon)}} \qquad \textit{where} \qquad \varepsilon = \frac{\log a_U^2}{\log a_{\mathrm{E}} + \log a_U^2}.$$

### FISTA convergence with a priori discretisation

We can summarise Lemmas 4.4.5 to 4.4.7 into a single FISTA iteration. The following theorem states the convergence guarantees when $\mathbb{U}^n$ and $n_k$ are chosen a priori.

**Theorem 4.4.8.** *Let $\{\widetilde{\mathbb{U}}^k \text{ s.t. } k \in \mathbb{N}\}$ be an $(a_U, a_{\mathrm{E}})$-discretisation for* E *and choose any $\mathbb{U}^n$ such that*

$$\widetilde{\mathbb{U}}^k = \mathbb{U}^{n_k}, \qquad \widetilde{w}_k \in \mathbb{U}^{n_k+1} \cap \ldots \cap \mathbb{U}^{n_{k+1}-1}$$

*for all* $k \in \mathbb{N}$. *Compute* $u_n$, $v_n$ *by Algorithm 4.1 and choose* $\widetilde{w}_k \in \operatorname{argmin}_{u\in\widetilde{\mathbb{U}}^k} \mathrm{E}(u)$.

*Suppose that either:*

- $a_U > 1$ *and* $n_k^2 \simeq a_{\mathrm{E}}^k a_U^{2k}$,
- *or* $a_U = 1$, $\sum_{k=1}^\infty n_k^2 a_{\mathrm{E}}^{-k} < \infty$ *and* $\sum_{k=1}^\infty \|\widetilde{w}_k - \widetilde{w}_{k+1}\| < \infty$,

*then*

$$\mathrm{E}_0(u_N) \lesssim \frac{1}{N^{2(1-\varepsilon)}} \qquad \textit{where} \qquad \varepsilon = \frac{\log a_U^2}{\log a_{\mathrm{E}} + \log a_U^2}$$

*uniformly for* $N \in \mathbb{N}$.

This theorem is very easy to implement and requires very little knowledge of how to estimate $\mathrm{E}_0(u_n)$. So long as $a_U$ and $a_{\mathrm{E}}$ can be computed analytically, choosing $\widetilde{\mathbb{U}}^k$ to be uniform discretisations and $\widetilde{\mathbb{U}}^k = \mathbb{U}^{n_k} = \ldots = \mathbb{U}^{n_{k+1}-1}$ will give the stated convergence rate.

### FISTA convergence with adaptivity

Lemma 4.4.6 gives a sufficient condition for converging at the rate $O(N^{2(\varepsilon-1)})$ but it is not necessary and in fact limits the potential convergence rate. To see this, if $a_{\mathrm{E}}$ is a sharp estimate of E, note that

$$\mathrm{E}_0(u_N) \geq \min_{u\in\mathbb{U}^N} \mathrm{E}_0(u) \simeq a_{\mathrm{E}}^{-K} \simeq \frac{a_U^{2K}}{N^2} \simeq N^{2(\varepsilon-1)}.$$

To go beyond this rate (for example the linear convergence which will be seen in Section 4.7.2) we need to consider choosing $n_k$ adaptively.

While Theorem 4.4.8 also allows for spatial adaptivity, we will make further comment here as it links strongly with the choice of $n_k$. Suppose $\{\widetilde{\mathbb{U}}^k \text{ s.t. } k \in \mathbb{N}\}$ is an $(a_U, a_{\mathrm{E}})$-discretisation for E. To ensure that $\{\mathbb{U}^{n_k} \text{ s.t. } k \in \mathbb{N}\}$ also satisfies the condition, it is sufficient to verify two properties:

$$u^{n_K-1} \in \mathbb{U}^{n_K} \subset \bigcup_{k=0}^K \widetilde{\mathbb{U}}^k \qquad \text{and} \qquad \mathrm{E}_0(u^{n_K-1}) \lesssim a_{\mathrm{E}}^{-K}.$$

The subspace inclusion guarantees the condition for $a_U$ and $\mathrm{E}_0(\widetilde{w}_K) = \min_{u\in\mathbb{U}^{n_K}} \mathrm{E}_0(u) \leq \mathrm{E}_0(u_{n_K-1})$ confirms the condition for $a_{\mathrm{E}}$.

For both adaptive choice of $n_k$ and $\mathbb{U}^n$, accurate estimation of $\mathrm{E}_0(u_n)$ is key. The idea of the following theorem is that we combine estimates of $\mathrm{E}_0(u_n)$ with a ‘backstop’ condition; ‘small’ refinements can happen at any time and ‘big’ refinements should happen as soon as $\mathrm{E}_0(u_n) \lesssim a_{\mathrm{E}}^{-k}$ for the appropriate $k \in \mathbb{N}$.

Without any way to estimate $\mathrm{E}_0(u_n)$, we must rely on Lemma 4.4.6 for a naive a priori bound. If $\mathrm{E}_0(u_n)$ can be computed exactly then we do not need to enforce $n_k^2 \lesssim a_{\mathrm{E}}^k a_U^{2k}$ at all; it is implicitly guaranteed. For any intermediate case, we should combine estimates and choose the best one. This is summarised by Theorem 4.4.9.

**Theorem 4.4.9.** *Let* $\{\mathbb{U}^n \subset \mathbb{H}\cap\mathbb{U} \text{ s. t. } n\in\mathbb{N}\}$ *be a sequence of subspaces and* $n_k\in\mathbb{N}$ *a monotone increasing sequence such that*

$$\widetilde{\mathbb{U}}^k := \mathbb{U}^{n_k} \ni u_{n_k-1}, \qquad \widetilde{w}_k \in \left[\operatorname*{argmin}_{u\in\widetilde{\mathbb{U}}^k} \mathrm{E}(u)\right] \cap \mathbb{U}^{n_k+1} \cap \ldots \cap \mathbb{U}^{n_{k+1}-1}$$

*for all* $k\in\mathbb{N}$*. Compute* $u_n$*,* $v_n$ *by Algorithm 4.1.*

*Suppose there exist* $a_U, a_\mathrm{E} \geq 1$ *such that either:*

- $a_U > 1$ *and* $n_k^2 \lesssim a_\mathrm{E}^k a_U^{2k}$*,*
- *or* $a_U = 1$*,* $\sum_{k=1}^\infty n_k^2 a_\mathrm{E}^{-k} < \infty$ *and* $\sum_{k=1}^\infty \|\widetilde{w}_k - \widetilde{w}_{k+1}\| < \infty$

*and both*

$$\|\widetilde{w}_k\| \lesssim a_U^k \qquad \text{and} \qquad \mathrm{E}_0(u_{n_K-1}) \lesssim a_\mathrm{E}^{-K}.$$

*Whenever these conditions on* $n_k$ *are satisfied, then*

$$\min_{n\leq N} \mathrm{E}_0(u_n) \lesssim \frac{1}{N^{2(1-\varepsilon)}} \qquad \text{where} \qquad \varepsilon = \frac{\log a_U^2}{\log a_\mathrm{E} + \log a_U^2}$$

*uniformly for* $N\in\mathbb{N}$*.*

This theorem can be directly compared with Theorem 4.4.8. In particular, we note that the convergence rate is the same in both theorems but the price for better adaptivity is a slightly weaker stability guarantee (i.e. the addition of the min on the left-hand side).

In a practical sense, Theorem 4.4.8 provides a worst case convergence rate but if the observed convergence is faster then the refinement should also be accelerated. The issue arises if E is a function which is ‘easy’ to minimise for small $n$ ($n_k$ is small) but ‘hard’ for large $n$ (Lemma 4.4.6 becomes sharp). If this behaviour is made to oscillate, then the convergence will also be oscillatory although still optimal in the sense of Theorem 4.4.9.

**Remark 4.4.10.** *The convergence guarantee of Theorem 4.4.9 can be strengthened back to the monotone statement* $\mathrm{E}_0(u_N) \lesssim \frac{1}{N^{2(1-\varepsilon)}}$ *whenever*

$$\widetilde{w}_k \in \bigcap_{n=n_k}^{N} \mathbb{U}^n$$

*for the appropriate* $k\in\mathbb{N}$*. This shows that the non-monotone convergence guarantee is somehow related to non-monotonicity of the discretisation, if* $\mathbb{U}^n\subset\mathbb{U}^{n+1}$ *for all* $n\in\mathbb{N}$ *then the above condition is always satisfied. It is currently unclear how to make this statement more precise without an excessive number of constraints on the choice of* $\mathbb{U}^n$*.*

Theorem 4.4.9 relies on estimation of when $\mathrm{E}_0(u_n) \lesssim a_\mathrm{E}^{-k}$ although there are many equivalent characterisations for this. Some are described in the following lemma.

**Lemma 4.4.11.** *Let* $\{\widetilde{\mathbb{U}}^k \text{ s.t. } k\in\mathbb{N}\}$ *be a sequence of subspaces with some points* $u_k\in\widetilde{\mathbb{U}}^k$ *and* $\widetilde{w}_k\in\operatorname{argmin}_{u\in\widetilde{\mathbb{U}}^k}\mathrm{E}(u)$*. Suppose that* $\|\widetilde{w}_k\|\lesssim a_U^k$*. Any of the following conditions are sufficient to show that* $\{\widetilde{\mathbb{U}}^k\}$ *is an* $(a_U, a_{\mathrm{E}})$*-discretisation for* $\mathrm{E}$*:*

1. *Small continuous gap refinement:* $\mathrm{E}_0(u_k)\leq\beta a_{\mathrm{E}}^{-k}$ *for all* $k\in\mathbb{N}$*, some* $\beta>0$*.*

2. *Small discrete gap refinement:* $\mathrm{E}_0(\widetilde{w}_k)\leq\beta a_{\mathrm{E}}^{-k}$ *and* $\mathrm{E}_0(u_k)-\mathrm{E}_0(\widetilde{w}_{k-1})\leq\beta a_{\mathrm{E}}^{-k}$ *for all* $k\in\mathbb{N}$*, some* $\beta>0$*.*

3. *Small relative gap refinement:* $\mathrm{E}_0(u_k)-\mathrm{E}_0(\widetilde{w}_{k-1})\leq\beta\,\mathrm{E}_0(u_k)$ *for all* $k\in\mathbb{N}$*, some* $0<\beta\leq\frac{1}{1+a_{\mathrm{E}}}$*.*

4. *Small continuous gradient refinement:* $|||\partial\,\mathrm{E}(u_k)|||_*\leq\beta a_{\mathrm{E}}^{-k}$ *for all* $k\in\mathbb{N}$*, some* $\beta>0$*, and sublevel sets of* $\mathrm{E}$ *are* $|||\cdot|||$*-bounded.*

5. *Small discrete gradient refinement:* $\mathrm{E}_0(\widetilde{w}_k)\leq\beta a_{\mathrm{E}}^{-k}$ *and* $|||\Pi_k\partial\,\mathrm{E}(u_k)|||_*\leq\beta a_{\mathrm{E}}^{-k}$ *for all* $k\in\mathbb{N}$*, some* $\beta>0$*, and sublevel sets of* $\mathrm{E}$ *are* $|||\cdot|||$*-bounded. The operator* $\Pi_k\colon\mathbb{H}\to\widetilde{\mathbb{U}}^k$ *is the orthogonal projection.*

The refinement criteria described by Lemma 4.4.11 can be split into two groups. Cases (1), (3), and (4) justify that $\widetilde{\mathbb{U}}^k$ is good enough whenever $u_k$ (or $u_{n_k-1}$ in Theorem 4.4.9) is contained in $\widetilde{\mathbb{U}}^k$. In cases (2) and (5), $u_k$ is good enough to choose $n_k$ but $u_k\in\widetilde{\mathbb{U}}^k$ is not sufficient to validate $\widetilde{\mathbb{U}}^k$ on its own.

Another splitting of the criteria is into gap and gradient computations. Typically, gradient norms (in (4) and (5)) should be easier to estimate than gaps because they only require local knowledge rather than global, i.e. $\partial\,\mathrm{E}(u_n)$ rather than an estimate of $\mathrm{E}(u^*)$. Implicitly, the global information comes from an extra condition on $\mathrm{E}$ to assert that sublevel sets are bounded.

## 4.5 General examples

In this work we consider the Lasso problem to be both a motivating example and source of numerical examples, however, Algorithm 4.1 is much more broadly applicable. The aim of this section is to justify this fact and demonstrate that the constants $a_U$ and $a_{\mathrm{E}}$ can be easily computed in many examples.

The trivial source of examples is any problem where $\mathbb{U}$ is finite dimensional. In this case Theorems 4.4.8 and 4.4.9 both achieve the standard FISTA rate, almost by default. The conditions on $\{\mathbb{U}^n \text{ s.t. } n\in\mathbb{N}\}$ prevent ‘oscillations’ in the discretisation and therefore Algorithm 4.1 becomes standard FISTA after a finite number of iterations (i.e. $n_k=\infty$ for some $k$).

For less trivial examples, we explore the use of finite element bases for discretisation when $\mathbb{H}=L^2(\Omega)$ for some compact $\Omega\subset\mathbb{R}^d$. Informally, we wish to generate spaces $\{\widetilde{\mathbb{U}}^k \text{ s.t. } k\in\mathbb{N}\}$ with the properties:

- each $\widetilde{\mathbb{U}}^k$ is a finite dimensional subspace of $L^\infty$,
- each $\widetilde{\mathbb{U}}^k$ can be spanned by shifting and rescaling elements of $\widetilde{\mathbb{U}}^0$,
- and $\widetilde{\mathbb{U}}^k$ are known to have good approximation rates in $|||\cdot|||$.

The first point makes it easy to compute the analytical equivalence between $\|\cdot\|$ and $|||\cdot|||$, the second point dictates the scaling with respect to $k$, the third point links back to optimisation. Our formal definitions for finite element spaces are given below.

**Definition 4.5.1.** *Suppose $\mathbb{U} \subset L^p(\Omega)$ for some compact domain $\Omega \subset \mathbb{R}^d$. We say that $\mathbb{M}^k = \{\omega_1^k, \omega_2^k, \ldots\}$ is a* mesh *if*

$$\omega_i^k \subset \Omega, \qquad |\omega_i^k \cap \omega_j^k| = 0$$

*for all $i$ and $j$. We say that $\widetilde{\mathbb{U}}^k \subset \mathbb{U} \cap \mathbb{H}$ is a* finite element space *if there exists a mesh $\mathbb{M}^k$ such that for all $u \in \widetilde{\mathbb{U}}^k$ there exists $u_i^k \in \widetilde{\mathbb{U}}^k$ such that*

$$\operatorname{supp}(u_i^k) \subset \overline{\omega_i^k}, \qquad u(\boldsymbol{x}) = u_i^k(\boldsymbol{x}) \qquad \text{for a.e. } \boldsymbol{x} \in \omega_i^k, \text{ all } i \in \mathbb{N}.$$

*We say that a sequence of finite element sub-spaces $\widetilde{\mathbb{U}}^k \subset \mathbb{U} \cap \mathbb{H}$ is* $h$-refining *if there exists a basis $\{e_1, \ldots, e_N\} \subset \widetilde{\mathbb{U}}^0$ such that for any $u_i^k \in \widetilde{\mathbb{U}}^k$ with $\operatorname{supp}(u_i^k) \subset \omega_i^k$ there exist $\alpha \in \mathbb{R}^{d\times d}$, $\boldsymbol{\beta} \in \mathbb{R}^d$, $\boldsymbol{\gamma} \in \mathbb{R}^N$ such that*

$$0 < \det(\alpha) \lesssim h^{-kd}, \quad u_i^k(\boldsymbol{x}) = \sum_{j=1}^N \gamma_j e_j(\alpha \boldsymbol{x} + \boldsymbol{\beta}) \quad \text{for a.e. } \boldsymbol{x} \in \omega_i^k, \quad \operatorname{supp}(e_j(\alpha \cdot + \boldsymbol{\beta})) \subset \overline{\omega_i^k}.$$

*We say that $(\widetilde{\mathbb{U}}^k)_k$ is* of order $p$ *if*

$$\min_{w \in \widetilde{\mathbb{U}}^k} |||w - u^*||| \lesssim_{u^*} h^{kp}.$$

*We allow the implicit constant to have any dependence on $u^*$ so long as it is finite. In the case of Sobolev spaces, we would expect an inequality of the form $\min_{w \in \widetilde{\mathbb{U}}^k} \|w - u^*\|_{W^{0,q}} \lesssim h^{kp} \|u^*\|_{W^{p,q}}$ (Strang, 1972).*

We note that any piecewise polynomial finite element space can be used to form a $h$-refining sequence of subspaces. Wavelets almost satisfy this definition but not the support condition, each space contains every scale of wavelet. Similarly, a Fourier basis does satisfy nice scaling properties but each basis vector has global support. Both of these exceptions are important and could be accounted for with further analysis but we focus on the more standard finite element case. The following theorem shows that all finite element spaces achieve the same basic rates of convergence in Algorithm 4.1, depending on the particular properties of $|||\cdot|||$ and E.

**Theorem 4.5.2.** *Suppose* $\mathbb{H} = L^2(\Omega)$ *for some compact domain* $\Omega \subset \mathbb{R}^d$ *and* $\|\cdot\|_q \lesssim |||\cdot|||$ *for some* $q \in [1, \infty]$*. If* $(\widetilde{\mathbb{U}}^k)_k$ *is a sequence of h-refining finite element spaces of order p then*

$$a_U \leq \left\{ \begin{array}{cc} 1 & q \geq 2 \\ \sqrt{h^{-d}} & q < 2 \end{array} \right. ,$$

$$a_{\mathrm{E}} \geq \begin{cases} h^{-p} & \mathrm{E} \ \textit{is } |||\cdot|||\textit{-Lipschitz at } u^* \\ h^{-2p} & \mathrm{E} \ \textit{is } |||\cdot|||\textit{-smooth at } u^* \end{cases} .$$

*If* $\mathrm{g}$ *is not Lipschitz at* $u^*$ *but* $\mathrm{f}$ *is* $|||\cdot|||$*-Lipschitz and*

$$\min_{w\in\widetilde{\mathbb{U}}^k} \{|||w - u^*||| \text{ s. t. } \mathrm{g}(w) \leq \mathrm{g}(u^*)\} \lesssim \min_{w\in\widetilde{\mathbb{U}}^k} |||w - u^*|||,$$

*then, again,* $a_{\mathrm{E}} \geq h^{-p}$*.*

The proof of this theorem is in Section C.2. The main take-home for this theorem is that the computation of $a_U$ and $a_{\mathrm{E}}$ is typically very simple and obvious given a particular choice of $|||\cdot|||$ and E. The only non-trivial case is when g is not smooth enough to estimate directly. So long as $\widetilde{\mathbb{U}}^k$ are sufficiently dense in sublevel sets of g, the smoothness of f is sufficient to give a convergence rate.

## 4.6 Lasso minimisation

We now return to the concrete example of Lasso which will be used for numerical results in Section 4.7. We consider three forms of Lasso which will be referred to as continuous, countable, and discete Lasso depending on whether the space $\mathbb{U}$ is $\mathcal{M}([0,1]^d)$, $\ell^1(\mathbb{R})$, or finite dimensional respectively. In each case, the energy can be written as

$$\mathrm{E}(u) = \tfrac{1}{2}\left\|\mathcal{A}u - \boldsymbol{\eta}\right\|_{\ell^2}^2 + \mu|||u||| \tag{4.5}$$

for some $\mathcal{A}\colon \mathbb{U} \cup \mathbb{H} \to \mathbb{R}^m$ and $\mu > 0$, where $|||\cdot||| = \|\cdot\|_1$.

The aim of this section is to develop all of the necessary tools for implementing Algorithm 4.1 on the energy (4.5) through either Theorem 4.4.8 or Theorem 4.4.9. This includes computing the rates $a_U$ and $a_{\mathrm{E}}$, estimating the continuous gap $\mathrm{E}_0(u_n)$, and developing an efficient refinement choice for $\mathbb{U}^n$.

### 4.6.1 Continuous case

We start by estimating rates in the case $\mathbb{U} = \mathcal{M}(\Omega)$ where $\Omega = [0,1]^d$. In this case we choose $\widetilde{\mathbb{U}}^k$ to be the span of all piecewise constant functions on a mesh of squares with maximum side length $2^{-k}$ (i.e. $h = {}^1\!/\!_2$).

We would like to apply Theorem 4.5.2 to compute the rate of convergence although $\|\cdot\|_1$ is too strong a metric and results in order $p=0$. This can be seen because for any $u\in L^1(\Omega^d)$,

$$\|u-\delta_0\|_1 = \sup_{\varphi\in C(\Omega^d), \|\varphi\|_\infty\leq 1} \langle \varphi,\ u-\delta_0\rangle = \|u\|_1 + \|\delta_0\|_1 \geq 1.$$

The result of Theorem 4.5.2 gives $a_U = 2^{\frac{d}{2}}$ (whether or not $p=0$) and choosing $|||\cdot|||$ to be an appropriate negative Sobolev norm or transport metric gives $p=1$ and $a_{\mathrm{E}} = 2^1$. We will work through this second statement explicitly to demonstrate the additional requirements on $\mathcal{A}$.

For any $w\in\widetilde{\mathbb{U}}^k$ such that $\|w\|_1\leq\|u^*\|_1$, we have

$$\begin{aligned}
\mathrm{E}(w)-\mathrm{E}(u^*) &= \tfrac{1}{2}\|\mathcal{A}w-\boldsymbol{\eta}\|^2 - \tfrac{1}{2}\|\mathcal{A}u^*-\boldsymbol{\eta}\|^2 + \mu\left(|||w||| - |||u^*|||\right)\\
&\leq \tfrac{1}{2}\|\mathcal{A}w-\boldsymbol{\eta}\|^2 - \tfrac{1}{2}\|\mathcal{A}u^*-\boldsymbol{\eta}\|^2\\
&= \left\langle \tfrac{1}{2}\mathcal{A}(w+u^*)-\boldsymbol{\eta},\ \mathcal{A}(w-u^*)\right\rangle\\
&\leq \max(\|\mathcal{A}w-\boldsymbol{\eta}\|, \|\mathcal{A}u^*-\boldsymbol{\eta}\|)\,\|\mathcal{A}(w-u^*)\|\\
&\leq \sqrt{2\,\mathrm{E}(w)}\,\|\mathcal{A}(w-u^*)\|.
\end{aligned}$$

For any bounded linear extension to the orthogonal projection $\widetilde{\Pi}_k\colon \mathbb{U}\to\widetilde{\mathbb{U}}^k$, choose $w=\widetilde{\Pi}_k u^*$, then we get

$$\mathrm{E}(w)-\mathrm{E}(u^*) \lesssim \left\|\mathcal{A}(\widetilde{\Pi}_k u^* - u^*)\right\| \leq \left\|(\widetilde{\Pi}_k - \mathrm{id})\mathcal{A}^*\right\|_{\ell^2\to L^\infty} |||u^*|||.$$

Expanding this operator norm further,

$$\begin{aligned}
\left\|(\widetilde{\Pi}_k-\mathrm{id})\mathcal{A}^*\right\|_{\ell^2\to L^\infty} = \sup_{\boldsymbol{r}\in\mathbb{R}^m}\max_{\boldsymbol{x}\in\Omega}\frac{|[\widetilde{\Pi}_k\mathcal{A}^*\boldsymbol{r}](\boldsymbol{x}) - [\mathcal{A}^*\boldsymbol{r}](\boldsymbol{x})|}{\|\boldsymbol{r}\|_{\ell^2}} &\leq \sup_{\boldsymbol{r}\in\mathbb{R}^m}\frac{\sqrt{d}2^{-k}\|\nabla[\mathcal{A}^*\boldsymbol{r}]\|_{L^\infty}}{\|\boldsymbol{r}\|_{\ell^2}}\\
&= \sqrt{d}2^{-k}|\mathcal{A}^*|_{\ell^2\to C^1}
\end{aligned}$$

where $\sqrt{d}2^{-k}$ is the diameter of a $d$ dimensional pixel of side length $2^{-k}$.

This computation confirms two things, firstly that the scaling constant is indeed $a_{\mathrm{E}}=2$, and secondly that the required smoothness to achieve a good rate with Algorithm 4.1 is that $\mathcal{A}^*\colon\mathbb{R}^m\to C^1(\Omega)$ is a bounded operator. This accounts for the fact that $\mathcal{M}(\Omega)$ is such a large space containing distributions. In Section 4.6.5 we will show two practical examples were this (semi)norm is accurately computable.

Inserting the computed rates into Theorem 4.4.8 or Theorem 4.4.9 gives the guaranteed convergence rate

$$\varepsilon = \frac{\log a_U^2}{\log a_{\mathrm{E}} + \log a_U^2} = \frac{d}{1+d} \qquad\Longrightarrow\qquad \mathrm{E}_0(u_N)\lesssim N^{-2(1-\varepsilon)} = N^{-\frac{2}{1+d}}. \tag{4.6}$$

This energy rate also corresponds to a resolution rate, on iteration $N$ with $N^2 \sim (a_{\mathrm{E}} a_U^2)^k$ we expect the resolution to be

$$h = 2^{-k} = \left(a_{\mathrm{E}} a_U^2\right)^{\frac{k}{1+d}} \sim N^{-\frac{2}{1+d}}. \tag{4.7}$$

### 4.6.2 Countable and discrete case

We now extend the rate computations to the case when $\mathbb{U} = \ell^1(\mathbb{R})$, or a finite dimensional subspace. The key fact here is that, even when $\mathbb{U}$ is infinite dimensional, it is known (e.g. (Unser et al., 2016, Theorem 2) and (Boyer et al., 2019, Corollary 2)) that $u^*$ can be chosen to have at most $m$ non-zeros. If this is the case, then $u^* \in \ell^2(\mathbb{R})$; this makes the analysis much simpler than in the continuous case. In this subsection we will ignore the discrete case as a simple sub-case of the countable.

For countable Lasso we consider discretisation subspaces of the form

$$\widetilde{\mathbb{U}}^k = \{u \in \ell^1(\mathbb{R}) \text{ s.t. } i \notin J_k \implies u_i = 0\}$$

for some sets $J_k \subset \mathbb{N}$, i.e. infinite vectors with finitely many non-zeros. The key change in analysis from the continuous case is that sparse vectors in $\ell^1(\mathbb{R})$ are also in $\ell^2(\mathbb{R})$. Because of this, $a_U = 1$ which leads to the expected optimal rate

$$\min_{n \leq N} \mathrm{E}_0(u_n) \lesssim \frac{1}{N^2},$$

independent of $a_{\mathrm{E}}$ or any additional properties of $\mathcal{A}$. The number of refinements will also be finite, therefore $n_k = \infty$ for some $k$ and the remaining conditions of Theorems 4.4.8 and 4.4.9 hold trivially.

### 4.6.3 Refinement metrics

Lemma 4.4.11 shows that adaptive refinement can be performed based on estimates of the function gap or the gradient. In this subsection we provide estimates for these values which can be easily computed.

#### Bounds for discretised functionals

We start by computing estimates for discretised Lasso. This covers the cases when either continuous/countable Lasso is projected onto $\mathbb{U}^n$, or $\mathbb{U}$ is finite dimensional. For notation we will use the continuous case. To recover the other cases, just replace continuous indexing ($u(\boldsymbol{x})$) with discrete ($u_i$).

Let $\Pi_n\colon \mathbb{U} \to \mathbb{U}^n$ denote the orthogonal projection (in fact the unique bounded extension). For Lasso, discretising the function E ($u \mapsto \mathrm{E}(\Pi_n u)$) is equivalent to replacing $u$ with $\Pi_n u$, and $\mathcal{A}^*$ with $\Pi_n\mathcal{A}^*$.

**Discrete gradient** We can use this formulation to compute the discrete sub-derivative at $u_n \in \mathbb{U}^n$:

$$\begin{aligned}\partial_n \mathrm{E}(\Pi_n u_n)(\boldsymbol{x}) &= [\Pi_n\mathcal{A}^*(\mathcal{A}u_n - \boldsymbol{\eta})](\boldsymbol{x}) + \begin{cases}\{+\mu\} & u_n(\boldsymbol{x}) > 0\\ [-\mu,\mu] & u_n(\boldsymbol{x}) = 0\\ \{-\mu\} & u_n(\boldsymbol{x}) < 0\end{cases}\\ &=: [\Pi_n\mathcal{A}^*(\mathcal{A}u_n - \boldsymbol{\eta})](\boldsymbol{x}) + \mu\Pi_n \operatorname{sign}(u_n(\boldsymbol{x}))\end{aligned}$$

where we define $s + \mu[-1,1] = [s-\mu, s+\mu]$ for all $s \in \mathbb{R}$, $\mu \geq 0$.

We need a metric to decide whether $\partial_n \mathrm{E}$ is small. As $|||\cdot||| = \|\cdot\|_1$, the natural metric is $|||\cdot|||_* = \|\cdot\|_\infty$ and so we get the estimate

$$\begin{aligned}|||\partial_n \mathrm{E}(u_n)|||_* &= \max_{\boldsymbol{x}\in\Omega} \min\left\{|v| \text{ s.t. } v \in \Pi_n\mathcal{A}^*(\mathcal{A}u_n - \boldsymbol{\eta})(\boldsymbol{x}) + \mu\operatorname{sign}(u_n(\boldsymbol{x}))\right\}\\ &= \max_{\boldsymbol{x}\in\Omega}\begin{cases}|[\Pi_n\mathcal{A}^*(\mathcal{A}u_n - \boldsymbol{\eta})(\boldsymbol{x}) + \mu| & u_n(\boldsymbol{x}) > 0\\ |[\Pi_n\mathcal{A}^*(\mathcal{A}u_n - \boldsymbol{\eta})(\boldsymbol{x}) - \mu| & u_n(\boldsymbol{x}) < 0\\ \max\left(|[\Pi_n\mathcal{A}^*(\mathcal{A}u_n - \boldsymbol{\eta})(\boldsymbol{x})| - \mu, 0\right) & u_n(\boldsymbol{x}) = 0\end{cases}\end{aligned}$$

which can be used in Lemma 4.4.11.

**Discrete gap** We now move on to the discrete gap, $\mathrm{E}(u_n) - \min_{u\in\mathbb{U}^n}\mathrm{E}(u)$. This can be computed with a dual representation, such as that used by Duval and Peyré (2017a),

$$\begin{aligned}\min_{u\in\mathbb{U}^n} \tfrac{1}{2}\|\mathcal{A}u - \boldsymbol{\eta}\|_{\ell^2}^2 + \mu|||u||| &= \min_{u\in\mathbb{H}}\max_{\boldsymbol{\varphi}\in\mathbb{R}^m}(\mathcal{A}\Pi_n u - \boldsymbol{\eta})\boldsymbol{\cdot}\boldsymbol{\varphi} + \mu|||\Pi_n u||| - \tfrac{1}{2}\|\boldsymbol{\varphi}\|_{\ell^2}^2\\ &= \max_{\boldsymbol{\varphi}\in\mathbb{R}^m}\min_{u\in\mathbb{H}}(\mathcal{A}\Pi_n u - \boldsymbol{\eta})\boldsymbol{\cdot}\boldsymbol{\varphi} + \mu|||\Pi_n u||| - \tfrac{1}{2}\|\boldsymbol{\varphi}\|_{\ell^2}^2\\ &= \max_{\boldsymbol{\varphi}\in\mathbb{R}^m}\begin{cases}-\boldsymbol{\eta}\boldsymbol{\cdot}\boldsymbol{\varphi} - \tfrac{1}{2}\|\boldsymbol{\varphi}\|_{\ell^2}^2 & |||\Pi_n\mathcal{A}^*\boldsymbol{\varphi}|||_* \leq \mu\\ -\infty & \text{else}\end{cases}\\ &= -\min_{\boldsymbol{\varphi}\in\mathbb{R}^m}\underbrace{\tfrac{1}{2}\|\boldsymbol{\varphi}\|_{\ell^2}^2 + \boldsymbol{\eta}\boldsymbol{\cdot}\boldsymbol{\varphi}}_{=:\mathrm{E}^*(\boldsymbol{\varphi})} + \chi(|||\Pi_n\mathcal{A}^*\boldsymbol{\varphi}|||_* \leq \mu).\end{aligned}$$

In particular,

$$\mathrm{E}(u) - \min_{u\in\mathbb{U}^n}\mathrm{E}(u) = \mathrm{E}(u) + \min_{|||\Pi_n\mathcal{A}^*\boldsymbol{\varphi}|||_*\leq\mu}\mathrm{E}^*(\boldsymbol{\varphi}) \leq \mathrm{E}(u) + \mathrm{E}^*(\boldsymbol{\varphi})$$

for any feasible $\boldsymbol{\varphi} \in \mathbb{R}^m$. Differentiating the saddle point with respect to $\boldsymbol{\varphi}$, if $\boldsymbol{\varphi}^*$ is the maximiser for $u^* \in \operatorname{argmin}_{u\in\mathbb{U}^n} \mathrm{E}(u)$ then

$$\boldsymbol{\varphi}^* = \mathcal{A}u^* - \boldsymbol{\eta}.$$

We remark briefly that it is more conventional to include the constraint in the definition of $\mathrm{E}^*$. We choose to omit it here to highlight that it is only the constraint which changes between the discrete and continuous cases; $\mathrm{E}^*$ will remain the same.

Given $u_n \in \mathbb{U}^n$, the optimality condition motivates a simple rule for choosing $\boldsymbol{\varphi}$:

$$\boldsymbol{\varphi}_n := \mathcal{A}u_n - \boldsymbol{\eta}, \qquad \mathrm{E}(u) - \min_{u'\in\mathbb{U}^n} \mathrm{E}(u') \leq \mathrm{E}(u) + \min_{\gamma\geq 0} \{\mathrm{E}^*(\gamma\boldsymbol{\varphi}_n) \text{ s.t. } \gamma|||\Pi_n\mathcal{A}^*\boldsymbol{\varphi}_n|||_* \leq \mu\}$$

with optimal choice

$$\gamma = \max\left(0, \min\left(\frac{-\boldsymbol{\eta}\cdot\boldsymbol{\varphi}_n}{\|\boldsymbol{\varphi}_n\|_{\ell^2}^2}, \frac{\mu}{|||\Pi_n\mathcal{A}^*\boldsymbol{\varphi}_n|||_*}\right)\right).$$

To apply Algorithm 4.1, we are assuming that both $\mathrm{f}(u_n) = \frac{1}{2}\|\boldsymbol{\varphi}_n\|_{\ell^2}^2$ and $\Pi_n\nabla\mathrm{f}(u_n) = \Pi_n\mathcal{A}^*\boldsymbol{\varphi}_n$ are easily computable, therefore $\gamma$ and $\mathrm{E}(u_n) + \mathrm{E}^*(\gamma\boldsymbol{\varphi}_n)$ are also easy to compute.

### Bounds for countable functionals

Extending the results of Section 4.6.3 to $\mathbb{U} = \ell^1(\mathbb{R})$ is analytically very simple but relies heavily on the specific choice of $\mathcal{A}$. The computations of gradients and gaps carry straight over replacing $\Pi_n$ with id and adding the sets $J_n \subset \mathbb{N}$ which define $\mathbb{U}^n = \{u \in \ell^1 \text{ s.t. } i \notin J_n \implies u_i = 0\}$. In particular,

$$|||\partial\mathrm{E}(u_n)|||_* = \max_{i\in\mathbb{N}} \begin{cases} |[\mathcal{A}^*\boldsymbol{\varphi}_n]_i + \mu| & [u_n]_i > 0 \\ |[\mathcal{A}^*\boldsymbol{\varphi}_n]_i - \mu| & [u_n]_i < 0 \\ \max(|[\mathcal{A}^*\boldsymbol{\varphi}_n]_i| - \mu, 0) & [u_n]_i = 0 \end{cases}$$

$$\mathrm{E}_0(u_n) \leq \mathrm{E}(u_n) + \mathrm{E}(\gamma_0\boldsymbol{\varphi}_n), \qquad \gamma_0 = \max\left(0, \min\left(\frac{-\boldsymbol{\eta}\cdot\boldsymbol{\varphi}_n}{\|\boldsymbol{\varphi}_n\|_{\ell^2}^2}, \frac{\mu}{|||\mathcal{A}^*\boldsymbol{\varphi}_n|||_*}\right)\right)$$

where $\boldsymbol{\varphi}_n = \mathcal{A}u_n - \boldsymbol{\eta} \in \mathbb{R}^m$ is always exactly computable.

In the countable case, the sets $J_n$ give a clear partition into known/unknown values in these definitions. For $i \in J_n$ the computation is the same as in Section 4.6.3, then for $i \notin J_n$ we know

$[u_n]_i = 0$ which simplifies the remaining computations. This leads to:

$$\begin{aligned}
|||\partial\,\mathrm{E}(u_n)|||_* &= \max\left(\max_{i\in J_n} |[\partial\,\mathrm{E}(u_n)]_i|,\ \max_{i\notin J_n} |[\partial\,\mathrm{E}(u_n)]_i|\right)\\
&= \max\left(|||\partial_n\,\mathrm{E}(u_n)|||_*,\ \max_{i\notin J_n} |[\mathcal{A}^*\boldsymbol{\varphi}_n]_i| - \mu\right)\\
|||\mathcal{A}^*\boldsymbol{\varphi}_n|||_* &= \max\left(\max_{i\in J_n} |[\mathcal{A}^*\boldsymbol{\varphi}_n]_i|,\ \max_{i\notin J_n} |[\mathcal{A}^*\boldsymbol{\varphi}_n]_i|\right)\\
&= \max\left(|||\Pi_n\mathcal{A}^*\boldsymbol{\varphi}_n|||_*,\ \max_{i\notin J_n} |[\mathcal{A}^*\boldsymbol{\varphi}_n]_i|\right).
\end{aligned}$$

Both estimates only rely on an upper bound of $\max_{i\notin J_n} |[\mathcal{A}^*\boldsymbol{\varphi}_n]_i|$. This must be computed on a per-example basis, one example is seen in Section 4.7.2.

### Bounds for continuous functionals

Finally, we extend the results of Section 4.6.3 to continuous Lasso. Similar to the countable case, the exact formulae can be written down immediately:

$$|||\partial\,\mathrm{E}(u_n)|||_* = \max_{\boldsymbol{x}\in\Omega}\begin{cases} |[\mathcal{A}^*\boldsymbol{\varphi}_n](\boldsymbol{x}) + \mu| & u_n(\boldsymbol{x}) > 0\\ |[\mathcal{A}^*\boldsymbol{\varphi}_n](\boldsymbol{x}) - \mu| & u_n(\boldsymbol{x}) < 0\\ \max\left(|[\mathcal{A}^*\boldsymbol{\varphi}_n](\boldsymbol{x})| - \mu, 0\right) & u_n(\boldsymbol{x}) = 0\end{cases}$$

$$\mathrm{E}_0(u_n) \leq \mathrm{E}(u_n) + \mathrm{E}(\gamma_0\boldsymbol{\varphi}_n), \qquad \gamma_0 = \max\left(0, \min\left(\frac{-\boldsymbol{\eta}\cdot\boldsymbol{\varphi}_n}{\|\boldsymbol{\varphi}_n\|_{\ell^2}^2}, \frac{\mu}{|||\mathcal{A}^*\boldsymbol{\varphi}_n|||_*}\right)\right).$$

If these quantities are not analytically tractable then we use the mesh $\mathbb{M}^n$ corresponding to $\mathbb{U}^n$ to decompose the bounds:

$$\begin{aligned}
|||\partial\,\mathrm{E}(u_n)|||_* &= \max_{\omega_i^n\in\mathbb{M}^n}\begin{cases} \|\mathcal{A}^*\boldsymbol{\varphi}_n + \mu\|_{L^\infty(\omega_i^n)} & u_n|_{\omega_i^n} > 0\\ \|\mathcal{A}^*\boldsymbol{\varphi}_n - \mu\|_{L^\infty(\omega_i^n)} & u_n|_{\omega_i^n} < 0\\ \max(0, \|\mathcal{A}^*\boldsymbol{\varphi}_n\|_{L^\infty(\omega_i^n)} - \mu) & u_n|_{\omega_i^n} = 0\end{cases}\\
|||\mathcal{A}^*\boldsymbol{\varphi}_n|||_* &= \max_{\omega_i^n\in\mathbb{M}^n} \|\mathcal{A}^*\boldsymbol{\varphi}_n\|_{L^\infty(\omega_i^n)}.
\end{aligned}$$

Now, both estimates rely on cell-wise supremum norms of $\mathcal{A}^*\boldsymbol{\varphi}_n$ which we assume is sufficiently smooth. We will use a cell-wise Taylor expansion to provide such an estimate which is both accurate and relatively tight. For instance, let $\boldsymbol{x}_i$ be the midpoint of the square $\omega_i^n$, then

$$\|\mathcal{A}^*\boldsymbol{\varphi}_n\|_{L^\infty(\omega_i^n)} \leq |[\mathcal{A}^*\boldsymbol{\varphi}_n](\boldsymbol{x}_i)| + \frac{\mathrm{diam}(\omega_i^n)}{2}|[\nabla\mathcal{A}^*\boldsymbol{\varphi}_n](\boldsymbol{x}_i)| + \frac{\mathrm{diam}(\omega_i^n)^2}{8}|\mathcal{A}^*\boldsymbol{\varphi}_n|_{C^2}.$$

In this work we chose a first order expansion because we are looking for extrema of $\mathcal{A}^*\boldsymbol{\varphi}_n$, i.e. we are most interested in the squares $\omega_i^n$ such that

$$|[\mathcal{A}^*\boldsymbol{\varphi}_n](\boldsymbol{x}_i)| \approx \mu, \qquad |[\nabla\mathcal{A}^*\boldsymbol{\varphi}_n](\boldsymbol{x}_i)| \approx 0, \qquad [\nabla^2\mathcal{A}^*\boldsymbol{\varphi}_n](\boldsymbol{x}_i) \preceq 0.$$

A zeroth order expansion would be optimally inefficient (approximating $|[\nabla\mathcal{A}^*\boldsymbol{\varphi}_n](\boldsymbol{x}_i)|$ with $|\mathcal{A}^*\boldsymbol{\varphi}_n|_{C^1}$) and a second order expansion would possibly be the most elegant but harder to implement. We found that a first order expansion was simple and efficient.

The bounds presented here for continuous Lasso emphasise the twinned properties required for adaptive mesh optimisation. The mesh should be refined greedily to the structures of $u^*$, but also must be sufficiently uniform to prove that $u^*$ is the function we are approximating. This is a classical exploitation/exploration trade-off; exploiting visible structure whilst searching for other structures which are not yet visible.

### 4.6.4 Support detection

The main motivation for using Lasso in applications is because it recovers sparse signals, in the case of compressed sensing the support of $u^*$ is also provably close to the 'true' support (Duval and Peyré, 2017a; Poon et al., 2018). If $u_n \approx u^*$ in the appropriate sense, then we should also be able to quantify the statement $\mathrm{supp}(u_n) \approx \mathrm{supp}(u^*)$. This is the aim of this subsection.

The work of Duval and Peyré (2017a); Poon et al. (2018) and many others characterise the support of $u^*$ very precisely. In particular, the support is at most $m$ distinct points and (with continuous notation) are a subset of $\{\boldsymbol{x} \in \Omega \text{ s.t. } |\mathcal{A}^*\boldsymbol{\varphi}^*|(\boldsymbol{x}) = \mu\}$. Less formally, this can also be seen from the the gradient computations in Section 4.6.3, for all $\boldsymbol{x} \in \mathrm{supp}(u^*)$

$$0 \in \partial\,\mathrm{E}(u^*)(\boldsymbol{x}) = [\mathcal{A}^*\boldsymbol{\varphi}^*](\boldsymbol{x}) + \mu\,\mathrm{sign}(u^*(\boldsymbol{x})).$$

From a computational perspective, identifying the support accurately allows for the most efficient choice of $\mathbb{U}^n$, however, to exclude areas from the support requires very accurate quantification.

Heuristically, we will use strong convexity of $\mathrm{E}^*$ and smoothness of $\mathcal{A}^*$ to quantify the statement:

$$\text{if} \quad \mathrm{E}(u_n) + \mathrm{E}^*(\gamma_0\boldsymbol{\varphi}_n) \approx 0 \quad \text{then} \quad \{\boldsymbol{x} \text{ s.t. } |[\mathcal{A}^*\boldsymbol{\varphi}_n](\boldsymbol{x})| \ll \mu\} \subset \{\boldsymbol{x} \text{ s.t. } u^*(\boldsymbol{x}) = 0\}.$$

First we use the strong convexity of $\mathrm{E}^*$, if $\gamma_0\boldsymbol{\varphi}_n$ and $\boldsymbol{\varphi}^*$ are both dual-feasible then

$$\tfrac{1}{2}\,\|\gamma_0\boldsymbol{\varphi}_n - \boldsymbol{\varphi}^*\|_{\ell^2}^2 \leq \mathrm{E}^*(\gamma_0\boldsymbol{\varphi}_n) - \mathrm{E}^*(\boldsymbol{\varphi}^*) = \mathrm{E}^*(\gamma_0\boldsymbol{\varphi}_n) + \mathrm{E}(u^*) \leq \mathrm{E}^*(\gamma_0\boldsymbol{\varphi}_n) + \mathrm{E}(u_n),$$

which gives an easily computable bound on $\|\gamma_0\boldsymbol{\varphi}_n - \boldsymbol{\varphi}^*\|_{\ell^2}$. Now we estimate $\mathcal{A}^*\boldsymbol{\varphi}_n$ on the support of $u^*$:

$$\begin{aligned}\min_{\boldsymbol{x}\in\operatorname{supp}(u^*)} |[\Pi_n\mathcal{A}^*\boldsymbol{\varphi}_n](\boldsymbol{x})| &\geq \min_{\boldsymbol{x}\in\operatorname{supp}(u^*)} |[\mathcal{A}^*\boldsymbol{\varphi}_n](\boldsymbol{x})| \\ &= \frac{1}{\gamma_0}\min_{\boldsymbol{x}\in\operatorname{supp}(u^*)} |[\mathcal{A}^*\gamma_0\boldsymbol{\varphi}_n](\boldsymbol{x})| \\ &\geq \frac{1}{\gamma_0}\min_{\boldsymbol{x}\in\operatorname{supp}(u^*)} |[\mathcal{A}^*\boldsymbol{\varphi}^*](\boldsymbol{x})| - |[\mathcal{A}^*\gamma_0\boldsymbol{\varphi}_n - \mathcal{A}^*\boldsymbol{\varphi}^*](\boldsymbol{x})| \\ &= \frac{1}{\gamma_0}\min_{\boldsymbol{x}\in\operatorname{supp}(u^*)} \mu - |[\mathcal{A}^*\gamma_0\boldsymbol{\varphi}_n - \mathcal{A}^*\boldsymbol{\varphi}^*](\boldsymbol{x})| \\ &\geq \frac{1}{\gamma_0}\left(\mu - |\mathcal{A}^*|_{\ell^2\to L^\infty}\|\gamma_0\boldsymbol{\varphi}_n - \boldsymbol{\varphi}^*\|_{\ell^2}\right).\end{aligned}$$

Therefore,

$$|[\mathcal{A}^*\gamma_0\boldsymbol{\varphi}_n](\boldsymbol{x})| < \mu - \sqrt{2(\mathrm{E}(u_n) + \mathrm{E}^*(\gamma_0\boldsymbol{\varphi}_n))}|\mathcal{A}^*|_{\ell^2\to L^\infty} \qquad \implies \qquad u^*(\boldsymbol{x}) = 0. \tag{4.8}$$

This equation is valid when $\boldsymbol{x}$ is either a continuous or countable index, the only distinction is to switch to $\ell^\infty$ in the norm of $\mathcal{A}^*$. To make the equivalent statement on the discretised problem, simply replace $\gamma_0$ with $\gamma$ and $\mathcal{A}^*$ with $\Pi_n\mathcal{A}^*$.

We can make two quick observations about this formula:

- The convergence guarantee from Theorem 4.4.8 is for the quantity $\mathrm{E}(u_n) - \mathrm{E}(u^*)$, the more relevant quantity here is $\mathrm{E}^*(\gamma_0\boldsymbol{\varphi}_n) + \mathrm{E}(u^*)$ for which there is no proven rate.
- In Section 4.6.1, $|\mathcal{A}^*|_{\ell^2\to C^1} < \infty$ was required to compute a rate of convergence, but only $|\mathcal{A}^*|_{\ell^2\to L^\infty} < \infty$ is needed to estimate the support.

### 4.6.5 Operator norms

For numerical implementation of Lasso, we are required to accurately estimate several operator norms of $\mathcal{A}$. For f to be 1-Lipschitz we must divide by $\|\mathcal{A}^*\mathcal{A}\|$, and the adaptivity described in Sections 4.6.1, 4.6.3 and 4.6.4 requires estimates of $|\mathcal{A}^*|_{\ell^2\to L^\infty}$, $|\mathcal{A}^*|_{\ell^2\to C^1}$, and $|\mathcal{A}^*|_{\ell^2\to C^2}$. The aim for this section is to provide estimates of these norms and seminorms for the numerical examples presented in Section 4.7.

We start by specifying the structure of $\mathcal{A}$. By linearity, there must exist kernels $\psi_j \in \mathbb{H}\cap\mathbb{U}^*$ such that $(\mathcal{A}u)_j = \langle\psi_j,\ u\rangle$ for all $u\in\mathbb{H}\cup\mathbb{U}$, $j = 1,\ldots,m$. In this form, the adjoint can be written precisely, $\mathcal{A}^*\colon \mathbb{R}^m \to \mathbb{H}$ by

$$[\mathcal{A}^*\boldsymbol{r}](\boldsymbol{x}) = \sum_{j=1}^m r_j\psi_j(\boldsymbol{x}) \qquad \text{for all } \boldsymbol{x}\in\Omega,\ \boldsymbol{r}\in\mathbb{R}^m.$$

The following lemma allows for exact computation of the operator norm of $\mathcal{A}$, as needed to ensure that f is 1-Lipschitz.

**Lemma 4.6.1.** *If $\mathcal{A}\colon \mathbb{H}\to\mathbb{R}^m$ has kernels $\psi_j\in\mathbb{H}$ for $j\in[m]$, then $\|\mathcal{A}^*\mathcal{A}\| = \|\mathcal{A}\mathcal{A}^*\|$ where $\mathcal{A}\mathcal{A}^*\in\mathbb{R}^{m\times m}$ has entries $(\mathcal{A}\mathcal{A}^*)_{i,j} = \langle\psi_i,\ \psi_j\rangle$ and the matrix norm is the standard spectral norm.*

*Proof.* The operator $\mathcal{A}$ has a finite dimensional range, therefore it also has a singular value decomposition. This shows that $\|\mathcal{A}^*\mathcal{A}\| = \|\mathcal{A}\mathcal{A}^*\|$. To compute the entries of $\mathcal{A}\mathcal{A}^*\colon\mathbb{R}^m\to\mathbb{R}^m$, observe that for any $\boldsymbol{r}\in\mathbb{R}^m$

$$(\mathcal{A}\mathcal{A}^*\boldsymbol{r})_i = \langle\psi_i,\ \mathcal{A}^*\boldsymbol{r}\rangle = \left\langle\psi_i,\ \sum_{j=1}^m r_j\psi_j\right\rangle = \sum_{j=1}^m \langle\psi_i,\ \psi_j\rangle\, r_j.$$

As required. □

If $\|\mathcal{A}^*\mathcal{A}\|$ is not analytically tractable, then Lemma 4.6.1 enables it to be computed using standard finite dimensional methods. The operator $\mathcal{A}\mathcal{A}^*$ is always finite dimensional and can be computed without discretisation error.

In the continuous case, when $\mathbb{H} = L^2(\Omega)$ we also need to estimate the smoothness properties of $\mathcal{A}^*$. A generic result for this is given in the following lemma.

**Lemma 4.6.2.** *If $\mathcal{A}\colon\mathbb{H}\to\mathbb{R}^m$ has kernels $\psi_j\in L^2(\Omega)\cap C^k(\Omega)$ for $j\in[m]$, then for all $\frac{1}{q}+\frac{1}{q^*}=1$, $q\in[1,\infty]$, we have*

$$|\mathcal{A}^*\boldsymbol{r}|_{C^k} \coloneqq \sup_{\boldsymbol{x}\in\Omega}|\nabla^k[\mathcal{A}^*\boldsymbol{r}]|(\boldsymbol{x}) \leq \sup_{\boldsymbol{x}\in\Omega}\left\|(\nabla^k\psi_j(\boldsymbol{x}))_{j=1}^m\right\|_{\ell^{q^*}}\|\boldsymbol{r}\|_{\ell^q},$$

$$|\mathcal{A}^*|_{\ell^2\to C^k} \coloneqq \sup_{\|\boldsymbol{r}\|_{\ell^2}\leq 1}|\mathcal{A}^*\boldsymbol{r}|_{C^k} \leq \sup_{\boldsymbol{x}\in\Omega}\left\|(\nabla^k\psi_j(\boldsymbol{x}))_{j=1}^m\right\|_{\ell^{q^*}}\times\begin{cases}1 & q\geq 2\\ \sqrt{m^{2-q}} & q<2\end{cases}.$$

*Proof.* For the first inequality, we apply the Hölder inequality on $\mathbb{R}^m$:

$$|\nabla^k[\mathcal{A}^*\boldsymbol{r}]|(\boldsymbol{x}) = \left|\sum_{j=1}^m\nabla^k\psi_j(\boldsymbol{x})r_j\right| \leq \left(\sum_{j=1}^m|\nabla^k\psi_j(\boldsymbol{x})|^{q^*}\right)^{\frac{1}{q^*}}\|\boldsymbol{r}\|_{\ell^q} = \left\|(\nabla^k\psi_j(\boldsymbol{x}))_j\right\|_{\ell^{q^*}}\|\boldsymbol{r}\|_{\ell^q}.$$

For the second inequality, if $q>2$ and $\sum_{j=1}^m r_j^2\leq 1$, then $|r_j|\leq 1$ for all $j$ and $\|\boldsymbol{r}\|_{\ell^q}^q\leq\|\boldsymbol{r}\|_{\ell^2}^2\leq 1$. If $q<2$ and $\|\boldsymbol{r}\|_{\ell^2}\leq 1$, then we again use Hölder's inequality:

$$\sum_{j=1}^m r_j^q \leq \left(\sum_{j=1}^m 1^{Q^*}\right)^{\frac{1}{Q^*}}\left(\sum_{j=1}^m r_j^{qQ}\right)^{\frac{1}{Q}} \leq m^{\frac{2-q}{2}}$$

for $Q=\frac{2}{q}$. □

### Concrete examples

Lemma 4.6.2 demonstrates how the smoothness of $\mathcal{A}$ relates to the smoothness of the kernels $\psi_j$, possibly scaling with respect to $m$. We now present explicit examples of the computations in both Lemmas 4.6.1 and 4.6.2 for the numerical results in Section 4.7. These examples are common in practical applications and we show the result of Theorem 4.6.3 in the main text to demonstrate that, while sometimes hard to compute by hand, the constants can estimated accurately in an efficient manner. The Gaussian case is by far the most technical, therefore we will provide a little further explanation at the end of the theorem.

**Theorem 4.6.3.** *Suppose $\mathcal{A}\colon \mathbb{H} \to \mathbb{R}^m$ has kernels $\psi_j \in \mathbb{H} = L^2([0,1]^d)$ for $j \in [m]$.*

*Case 1: If $\psi_j(\boldsymbol{x}) = \begin{cases} 1 & \boldsymbol{x} \in \mathbb{X}_j \\ 0 & else \end{cases}$ for some collection $\mathbb{X}_j \subset \Omega$ such that $\mathbb{X}_i \cap \mathbb{X}_j = \emptyset$ for all $i \neq j$, then*

$$\|\mathcal{A}\|_{L^2\to\ell^2} = \max_j \sqrt{|\mathbb{X}_j|}.$$

*Case 2: If $\psi_j(\boldsymbol{x}) = \cos(\boldsymbol{a}_j \cdot \boldsymbol{x})$ for some frequencies $\boldsymbol{a}_j \in \mathbb{R}^d$ with $|\boldsymbol{a}_j| \leq A$, then*

$$\|\mathcal{A}\|_{L^2\to\ell^2} \leq \sqrt{m}, \qquad |\mathcal{A}^*\boldsymbol{r}|_{C^k} \leq m^{1-\frac{1}{q}} A^k \left\|\boldsymbol{r}\right\|_q, \qquad |\mathcal{A}^*|_{\ell^2\to C^k} \leq \sqrt{m}A^k$$

*for all $\boldsymbol{r} \in \mathbb{R}^m$ and $q \in [1,\infty]$.*

*Case 3: Suppose $\psi_j(\boldsymbol{x}) = (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}_j|^2}{2\sigma^2}\right)$ for some regular mesh $\boldsymbol{x}_j \in [0,1]^d$ and separation $\Delta$. i.e.*

$$\{\boldsymbol{x}_j \text{ s.t. } j \in [m]\} = \{\boldsymbol{x}_0 + (j_1\Delta, \ldots, j_d\Delta) \text{ s.t. } j_i \in [\widehat{m}]\}$$

*for some $\boldsymbol{x}_0 \in \mathbb{R}^d$, $\widehat{m} := \sqrt[d]{m}$. For all $\frac{1}{q} + \frac{1}{q^*} = 1$, $q \in (1,\infty]$, we get*

$$\|\mathcal{A}\|_{L^2\to\ell^2} \leq \left( (4\pi\sigma^2)^{-\frac{d}{2}} \sum_{j=-2\widehat{m},\ldots,2\widehat{m}} \exp(-\tfrac{\Delta^2}{4\sigma^2}j^2)\right)^d,$$

$$|\mathcal{A}^*\boldsymbol{r}|_{C^0} \leq (2\pi\sigma^2)^{-\frac{d}{2}} \left( \sum_{\boldsymbol{j}\in J} \exp\left(-\tfrac{q^*\Delta^2}{2\sigma^2}\max(0,|\boldsymbol{j}|-\delta)^2\right)\right)^{\frac{1}{q^*}} \left\|\boldsymbol{r}\right\|_q,$$

$$|\mathcal{A}^*\boldsymbol{r}|_{C^1} \leq \frac{(2\pi\sigma^2)^{-\frac{d}{2}}}{\sigma}\frac{\Delta}{\sigma} \left( \sum_{\boldsymbol{j}\in J} (|\boldsymbol{j}|+\delta)^{q^*} \exp\left(-\tfrac{q^*\Delta^2}{2\sigma^2}\max(0,|\boldsymbol{j}|-\delta)^2\right)\right)^{\frac{1}{q^*}} \left\|\boldsymbol{r}\right\|_q,$$

$$|\mathcal{A}^*\boldsymbol{r}|_{C^2} \leq \frac{(2\pi\sigma^2)^{-\frac{d}{2}}}{\sigma^2} \left( \sum_{\boldsymbol{j}\in J} \left(1+\tfrac{\Delta^2}{\sigma^2}(|\boldsymbol{j}|+\delta)^2\right)^{q^*} \exp\left(-\tfrac{q^*\Delta^2}{2\sigma^2}\max(0,|\boldsymbol{j}|-\delta)^2\right)\right)^{\frac{1}{q^*}} \left\|\boldsymbol{r}\right\|_q,$$

*where $\delta = \frac{\sqrt{d}}{2}$ and $J = \{\boldsymbol{j} \in \mathbb{Z}^d \text{ s.t. } \|\boldsymbol{j}\|_{\ell^\infty} \leq 2\widehat{m}\}$. The case for $q = 1$ can be inferred from the standard limit of $\left\|\cdot\right\|_{q^*} \to \left\|\cdot\right\|_\infty$ for $q^* \to \infty$. For $\Delta \ll \sigma$ (i.e. high resolution* *data), we get the scaling behaviour*

$$\|\mathcal{A}\|_{L^2\to\ell^2} \lesssim \Delta^{-d}, \qquad |\mathcal{A}^*\boldsymbol{r}|_{C^k} \lesssim \sigma^{-k}\Delta^{-\frac{d}{q^*}}\|\boldsymbol{r}\|_q, \qquad |\mathcal{A}^*|_{\ell^2\to C^k} \lesssim \sigma^{-k}\Delta^{-\frac{d}{2}},$$

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

The sharpest method for estimating these norms is of course to compute them. This is most cumbersome for Gaussian kernels, however, the sums converge faster than exponentially and so the computational burden should be very small. On the other hand, if $\Delta$ is small, then the sum is just a quadrature estimate for some continuous integrals. These integrals can be computed analytically and provide a much more intuitive grasp of the scaling with respect to dimensionality, grid spacing, and Gaussian width.

## 4.7 Numerical examples

We present four numerical Lasso examples. The first two are in 1D to demonstrate the performance of different variants of Algorithm 4.1, both with and without adaptivity. In particular, we explore sparse Gaussian deconvolution and sparse signal recovery from Fourier data; each displays slightly different behaviour in terms of optimisation. We compare with the *continuous basis pursuit* (CBP) discretisation (Ekanadham et al., 2011; Duval and Peyré, 2017b) which is also designed to achieve super-resolution accuracy within a convex framework. More details of this method will be provided in Section 4.7.1.

The next example is 2D reconstruction from Radon or X-ray data with wavelet-sparsity. As the forward operator is not sufficiently smooth, we must optimise in $\ell^1(\mathbb{R})$ which naturally leads to the choice of a wavelet basis.

Finally, we process a dataset which represents a realistic application and (simulated) dataset for Algorithm 4.1 in biological microscopy, referred to as STORM microscopy. In essence, the task is to perform 2D Gaussian de-blurring/super-resolution and denoising to find the location of sparse spikes of signal.

In this section, the main aim is to minimise the exact Lasso energy $\mathrm{E}_0(u_n)$ and so this will be our main metric for the success of an algorithm, referred to as the 'continuous gap'. Lemma 4.4.11 only provides guarantees on the values of $\min_{n\leq N}\mathrm{E}_0(u_n)$ so it is this monotone estimate which is plotted. To clarify, as $\mathrm{E}(u^*)$ is not known exactly, we always use the estimate $\min_{n\leq N}\mathrm{E}_0(u_n) \approx \min_{n\leq N}\mathrm{E}(u_n) + \min_{n'\leq n}\mathrm{E}^*(\gamma_0\boldsymbol{\varphi}_{n'})$. Another quantity of interest is minimisation of the discrete Lasso energy $\min_{n\leq N}\mathrm{E}(u_n) + \min_{n'\leq n}\mathrm{E}^*(\gamma\boldsymbol{\varphi}_{n'})$ which will be referred to as the 'discrete gap'. Note that for the adaptive schemes, this may not be exactly monotonic because the discrete dual problem is changing with $N$.

### 4.7.1 1D continuous Lasso

In this example we choose $\mathbb{U}=\mathcal{M}([0,1])$ and $\mathcal{A}\colon \mathbb{U}\to\mathbb{R}^{30}$ with either random Fourier kernels:

$$(\mathcal{A}u)_j=\int_0^1 \cos(a_jx)u(x), \qquad a_j\sim \mathrm{Uniform}[-100,100],\ j\in[30],\ \mu=0.02,$$

or Gaussian kernels on a regular grid:

$$(\mathcal{A}u)_j=(2\pi\sigma^2)^{-\frac{1}{2}}\int_0^1 \exp\left(-\frac{(x-(j-1)\Delta)^2}{2\sigma^2}\right)u(x), \quad \sigma=0.12,\ \Delta=\tfrac{1}{29},\ j\in[30],\ \mu=0.06.$$

Many variants of FISTA are compared for these examples but the key alternative viewed here is the CBP discretisation. There are many methods designed to minimise the continuous Lasso energy in (4.5) (Bredies and Pikkarainen, 2013; De Castro et al., 2016; Boyd et al., 2017; Catala et al., 2019) however, most result in a non-convex discretised problem to solve. We have focused on CBP because it approximates $u^*$ through a convex discrete optimisation problem which is asymptotically exact in the limit $h\to 0$. It can even be solved efficiently with FISTA which allows for direct comparison with the uniform and adaptive mesh approaches. As explained by Ekanadham et al. (2011); Duval and Peyré (2017b), the idea is that for a fixed mesh, the kernels of $\mathcal{A}$ are expanded to first order on each pixel and a particular first order basis is also chosen. If $u^*$ has only one Dirac spike in each pixel then the zeroth order information should correspond to the mass of the spike and additional first order information should determine the location.

As shown in Section 4.6, in 1D we have $a_U=a_{\mathrm{E}}=2$. The estimates given in (4.6) and (4.7) in dimension $d=1$ predict that the adaptive energy will decay at a rate of $\mathrm{E}_0(u_n)\lesssim\frac{1}{n}$ so long as the pixel size also decreases at a rate of $h\sim\frac{1}{n}$. To achieve these rates, we implement a refinement criterion from Lemma 4.4.11 with guarantee of $\mathrm{E}_0(u_{n_k-1})\lesssim 2^{-k}$ using the estimates made in Section 4.6.3. We choose subspaces $\mathbb{U}^n$ to approximately enforce

$$\mathrm{E}(u_n)+\mathrm{E}^*(\gamma_0\boldsymbol{\varphi}_n)\leq 2(\mathrm{E}(u_n)+\mathrm{E}^*(\gamma\boldsymbol{\varphi}_n))$$

i.e. the continuous gap is bounded by twice the discrete gap. In particular, note that for $\gamma_0\approx\gamma$,

$$\mathrm{E}^*(\gamma_0\boldsymbol{\varphi}_n)=\tfrac{1}{2}\left\|\gamma_0\boldsymbol{\varphi}_n\right\|^2+\gamma_0\boldsymbol{\eta}\boldsymbol{\cdot}\boldsymbol{\varphi}_n=\frac{\gamma_0}{\gamma}\left(\frac{\gamma_0}{\gamma}\tfrac{1}{2}\left\|\gamma\boldsymbol{\varphi}_n\right\|^2+\gamma\boldsymbol{\eta}\boldsymbol{\cdot}\boldsymbol{\varphi}_n\right)\approx\frac{\gamma_0}{\gamma}\,\mathrm{E}^*(\gamma\boldsymbol{\varphi}_n).$$

Inserting this back into the continuous/discrete gap inequality, it becomes

$$\frac{\gamma_0}{\gamma}\leq 1+(2-1)\frac{\mathrm{E}(u_n)+\mathrm{E}^*(\gamma\boldsymbol{\varphi}_n)}{\mathrm{E}^*(\gamma\boldsymbol{\varphi}_n)}$$

where 2 was our initial chosen ratio. Converting this into a spatial refinement criteria, recall

$$\frac{\gamma_0}{\gamma} \approx \frac{|||\mathcal{A}^*\varphi_n|||_*}{|||\Pi_n\mathcal{A}^*\varphi_n|||_*} = \frac{\max_{\omega_i^n\in\mathbb{M}^n}\|\mathcal{A}^*\varphi_n\|_{L^\infty(\omega_i^n)}}{\max_{\omega_i^n\in\mathbb{M}^n}|\Pi_n\mathcal{A}^*\varphi_n(\omega_i^n)|} \approx \max_{\omega_i^n\in\mathbb{M}^n}\frac{\|\mathcal{A}^*\varphi_n\|_{L^\infty(\omega_i^n)}}{|\Pi_n\mathcal{A}^*\varphi_n(\omega_i^n)|}.$$

If $\frac{\gamma_0}{\gamma}$ is large then there must be pixels (values of $i$) in which this ratio is large. Because of the smoothness of $\mathcal{A}^*\varphi_n$, refining these pixels will reduce the halve the difference and reduce the balance again. This was found to be an efficient method of selecting pixels for refinement based on quantities which had already been computed. Note briefly that from a heuristic standpoint we are refining for two reasons, aligning with an exploitation/exploration viewpoint. Either the Taylor expansion is tight and it is likely that $u^*$ has support in this point, or the Taylor expansion is loose and we refine to improve our level of uncertainty. This second point guarantees that we find the whole support of $u^*$, not just a subset.

**Comparison of discretisation methods** In Figure 4.1 we compare the three core approaches: fixed uniform discretisation, adaptive discretisation, and CBP. In particular, we wish to observe their convergence properties as the number of pixels is allowed to grow. In each case we use a FISTA stepsize of $t_n = \frac{n+19}{20}$. The adaptive discretisation is started with one pixel and limited to 128, 256, or 512 pixels while the fixed and CBP discretisations have uniform discretisations with the maximum number of pixels. We observe:

- Increasing the number of pixels always increases the accuracy of the optimisation.
- The adaptive scheme is much more efficient, in both examples the adaptive scheme with 128 pixels is at least competitive with both fixed discretisations with 512 pixels. In fact, only a maximum of 214 pixels were needed by the adaptive method.
- With Fourier kernels the uniform piecewise constant discretisation is more efficient than CBP but in the Gaussian case this is reversed. This suggests that the CBP *does* achieve super-resolution when $\mathcal{A}$ is sufficiently smooth but may be less accurate when the kernels oscillate on the length-scale of a single pixel.
- The discrete gaps for non-adaptive optimisation behave as is common for FISTA, initial convergence is polynomial until a locally linear regime activates (Tao et al., 2016). CBP is always slower to converge than the piecewise constant discretisation.
- For the adaptive method, the continuous/discrete gaps are very similar for all $n$, as enforced by the refinement criterion.

It is not completely fair to judge CBP with the continuous gap because, although it generates a continuous representation, this continuous representation is not necessarily consistent with the discrete gap being optimised, unlike when discretised with finite element methods. On the other hand, this is still the intended interpretation of the algorithm and we have no more appropriate metric in this case.

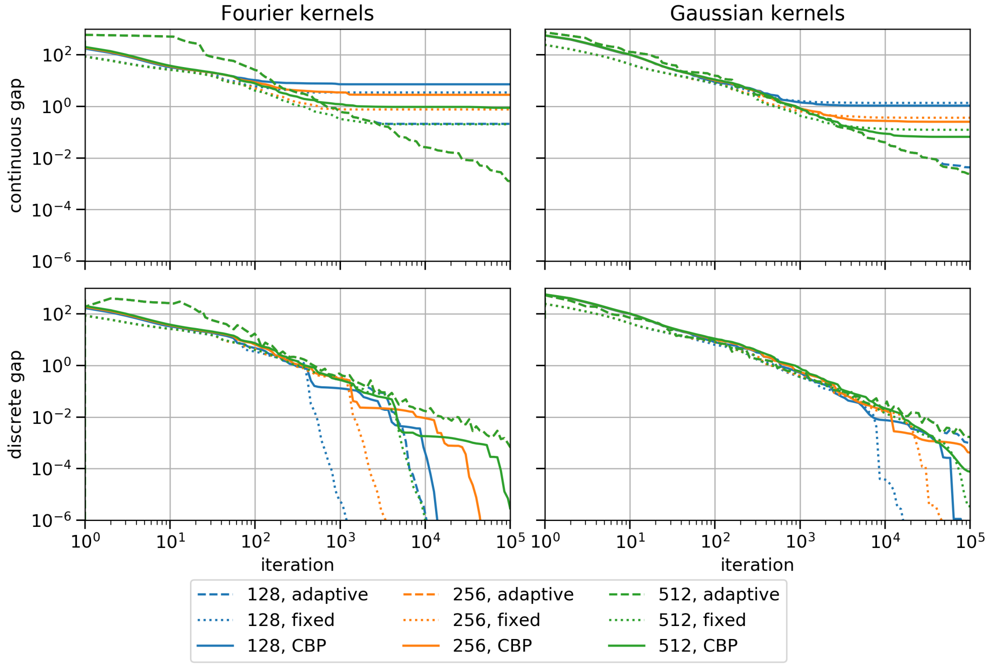


**Figure 4.1** Rates of continuous/discrete gap convergence for different Lasso algorithms with 128, 256, or 512 pixels. The ‘adaptive’ method uses the proposed algorithm. Both ‘fixed’ and ‘CBP’ use standard FISTA with a uniform discretisation.

**Comparison of FISTA variants** There are many variants of FISTA which can also be implemented in the form of Algorithm 4.1 just by updating the $\mathbb{U}^n$ on each iteration. Although we have not proven convergence for all of these, Figure 4.2 compares many methods with either fixed or adaptive discretisations. Each adaptive scheme is allowed up to 1024 pixels and each uniform discretisation uses exactly 1024. Forward-Backward splitting (FB) uses a sequence $t_n = 1$ otherwise for FISTA a general $t_n = \frac{n+a-1}{a}$ is used. The restarting scheme is given in Algorithm 1.2 and ‘greedy’ FISTA is given in Algorithm 1.3. In this example CBP used the greedy FISTA implementation which gave faster observed convergence. Figure 4.2 compares the discrete gaps because it is the accurate metric for fixed discretisations, and for the adaptive discretisation it should also be an accurate predictor of the continuous gap.

The key observations are:

- The only algorithm with noticeably different convergence is FB, which is the non-accelerated form of FISTA. Every other algorithm converges at the same approximate rate.
- The fixed discretisation schemes have an initial 'slow' convergence before reaching a 'fast' rate. The solid green line of FISTA $a = 2$ appears to achieve the theoretical $\frac{1}{n^2}$ rate and other FISTA implementations are much faster for large $n$.
- During the initial 'slow' phase, adaptive and fixed discretisations appear to achieve very similar (discrete) convergence rates. The coarse-to-fine adaptivity is not slower than fixed discretisations in this regime.
- Lemma 4.4.11 accurately predicts the $\frac{1}{n}$ rate of the adaptive methods, mirrored in the fixed discretisations. This suggests that high-resolution but fixed discretisations are initially limited by the continuous problem before entering the asymptotic discrete regime.
- Lemma 4.4.11 only applies to two adaptive schemes, labelled $a = 2$ and $a = 20$. The remaining FISTA schemes all perform comparably although the restarting scheme is often the slowest. Both $a = 20$ and greedy FISTA are consistently the best or near-best performing methods.

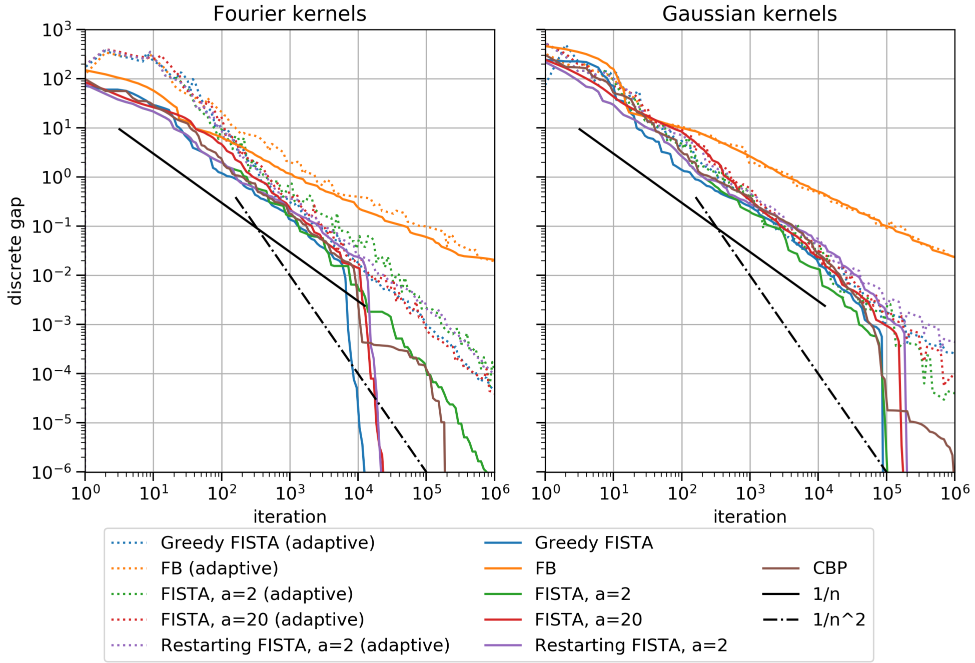


**Figure 4.2** Discrete convergence of different algorithms. 'Adaptive' methods use Algorithm 4.1 with fewer than 1024 pixels and the remaining methods use a uniform discretisation of 1024 pixels.

**Comparison of fixed and adaptive discretisation** Motivated by the findings in Figure 4.2, we now look more closely at the performance of the $a = 20$ and the greedy FISTA schemes. We have analytical results for the former but the latter typically performs the best for non-adaptive optimisation and is never worse than $a = 20$ in the adaptive setting. We assume that the aim is to find a function $u_n$ with $\mathrm{E}_0(u_n)$ smaller than a given threshold. The question is whether it is faster/more efficient to use the proposed adaptive scheme or to use a classical scheme at sufficiently high uniform resolution. The fixed discretisations use 1024 pixels (i.e. uniform pixel size of $2^{-10}$) and the adaptive discretisation starts with two pixels with an upper limit of 1024.

Figure 4.3 shows the convergence with respect to number of iterations. As expected, the fixed discretisation starts with a smaller continuous gap before plateauing to a sub-optimal gap around $\mathrm{E}_0 = 0.1$. In both examples, the greedy FISTA has much faster convergence around $n = 100$ and of course the minimum pixel size is constant for the fixed discretisation.

The adaptive optimisation matches the predicted rates well, both gap and minimum pixel size (equal to $2^{-k}$) decay at a rate of approximately $\frac{1}{n}$. Interestingly, in the Fourier case the energy decays a little faster and the resolution is a little slower. This is consistent with E being a little bit smoother than predicted (i.e. $\gamma_{\mathrm{E}} > 2$).

It is clear that the adaptive scheme is able to continue reducing the continuous gap far beyond that of the fixed discretisation. The range $n \in [10^3, 10^4]$ is particularly interesting because it is the time when the adaptive and fixed curves intersect in both continuous gap and minimum pixel size. Suppose the stopping criterion is to find $u$ such that $\mathrm{E}_0(u) < 0.1$. Figure 4.3 shows that it is equivalent to 'guess' the necessary resolution, or to adaptively refine until reaching the stopping criterion. Both methods would converge after $O(10^3)$ iterations with a minimum pixel size of $2^{-10}$.

Figure 4.4 shows a more practical comparison showing wall-clock computation time and number of pixels (memory usage). Optimisation of the fixed discretisation is faster overall but after around 0.1 s, it is always faster to use the adaptive scheme to achieve a given continuous gap. The reason for this can be seen in the numbers of pixels. At the most extreme, in the same computation time the adaptive scheme can achieve more than a factor of 10 better gap using approximately a factor of 10 fewer pixels. The adaptive scheme re-computes the discrete matrix $\mathcal{A}\Pi_n$ each time there is a refinement, but the fixed schemes only compute it once. The adaptive schemes still seem to converge faster than $\frac{1}{\mathrm{time}}$.

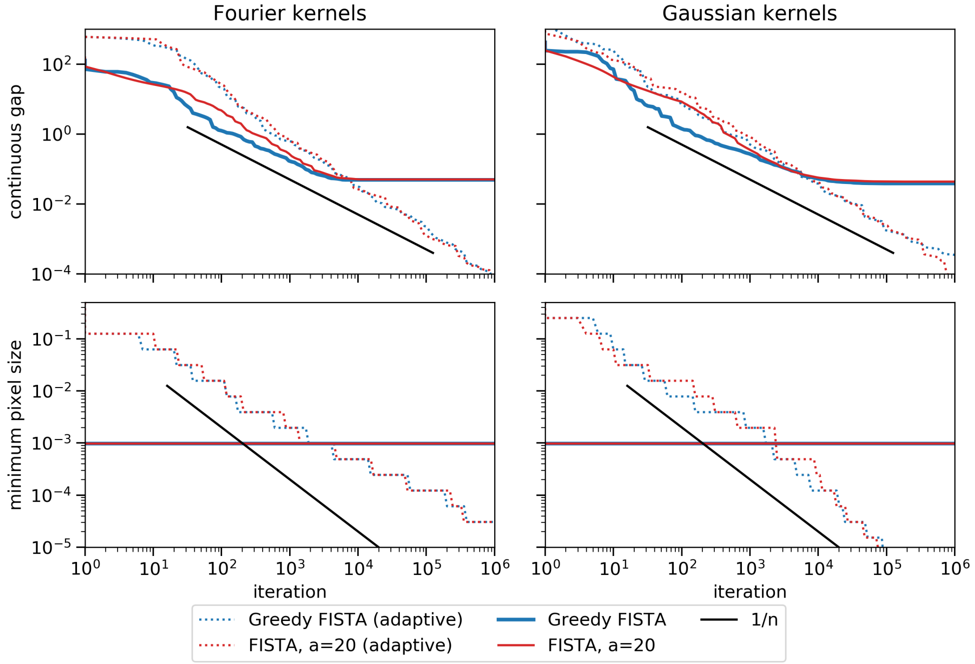


**Figure 4.3** Continuous convergence of adaptive (coarse-to-fine pixel size) compared with uniform discretisation (constant pixel size) with respect to number of iterations.

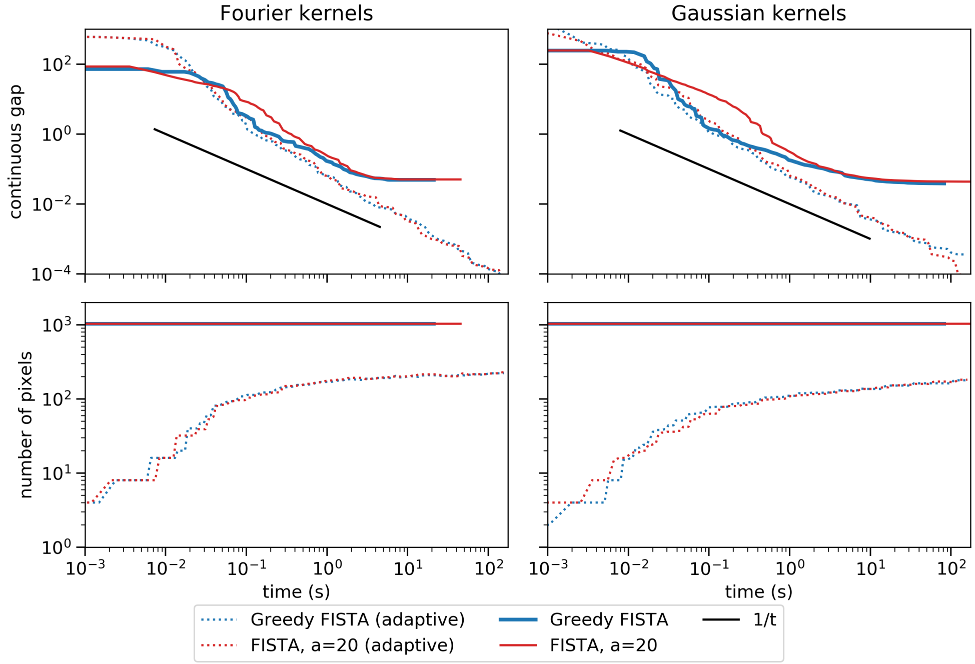


**Figure 4.4** Continuous convergence of adaptive compared with uniform discretisation with respect to wall-clock time and total number of pixels (memory requirement).

### 4.7.2 2D wavelet Lasso

In this example we consider $\mathcal{A}$ to be a 2D Radon transform. In particular, the rows of $\mathcal{A}$ correspond to integrals over the sets $\mathbb{X}_i^I$ where

$$\mathbb{X}_i^I = \left\{\boldsymbol{x} \in [-\tfrac12, \tfrac12]^2 \text{ s.t. } \boldsymbol{x}\boldsymbol{\cdot}\begin{pmatrix}\cos\theta_I\\ \sin\theta_I\end{pmatrix} \in \left[-\tfrac12 + \tfrac{i-1}{50}, -\tfrac12 + \tfrac{i}{50}\right)\right\}, \qquad \theta_I = \tfrac{180°}{51}I \qquad \text{for } i, I \in [50].$$

This is not exactly in the form analysed by Theorem 4.6.3, however for each $I$ the sets $\{\mathbb{X}_i^I \text{ s.t. } i \in [50]\}$ are disjoint therefore we can apply Theorem 4.6.3 block-wise to estimate

$$\|\mathcal{A}\|_{L^2\to\ell^2} \leq \sqrt{\sum_{I\in[50]} \max_{i\in[50]} |\mathbb{X}_i^I|} = \sqrt{\sum_{I\in[50]} \max_{i\in[50]} \int_{\mathbb{X}_i^I} 1 d\boldsymbol{x}} = \sqrt{\sum_{I\in[50]} \max_{i\in[50]} (\mathcal{A}\mathbb{1})_{i,I}}.$$

$\mathcal{A}$ is not smooth, therefore we can't bound $|\mathcal{A}^*|_{C^k}$ for $k > 0$, and so we must look to minimise over $\ell^1$ rather than $L^1$. The natural choice is to promote sparsity in a wavelet basis which can be rearranged into the Lasso form:

$$\min_{u\in\mathbb{U}} \tfrac12 \|\mathcal{A}u - \boldsymbol{\eta}\|_{\ell^2}^2 + \mu \left\|\mathcal{W}^{-1}u\right\|_{\ell^1} = \min_{\widehat{u}\in\ell^1(\mathbb{R})} \tfrac12 \|\mathcal{A}\mathcal{W}\widehat{u} - \boldsymbol{\eta}\|_{\ell^2}^2 + \mu \|\widehat{u}\|_{\ell^1}.$$

The minimisers are related by $u^* = \mathcal{W}\widehat{u}^*$ and, for wavelet bases, $\mathcal{W}$ is orthonormal so $\|\mathcal{A}\mathcal{W}\|_{\ell^2\to\ell^2} = \|\mathcal{A}\|_{L^2\to\ell^2}$. From Section 4.6.3 we know that to track convergence and perform adaptive refinement, it is sufficient to accurately bound $|[\mathcal{W}^\top\mathcal{A}^*\boldsymbol{\varphi}_n]_j|$ for all $j \notin J_n$. If $\mathcal{W}$ is a wavelet transformation then its columns, $w_j \in L^2$, are simply the wavelets themselves and we can use the bound

$$|\langle w_j,\ \mathcal{A}^*\boldsymbol{\varphi}_n\rangle| = \left|\left\langle w_j,\ \mathbb{1}_{\operatorname{supp}(w_j)}\mathcal{A}^*\boldsymbol{\varphi}_n\right\rangle\right| \leq \left\|\mathbb{1}_{\operatorname{supp}(w_j)}\mathcal{A}^*\boldsymbol{\varphi}_n\right\|_{L^2} \leq \|\mathbb{1}_{\mathbb{X}}\mathcal{A}^*\boldsymbol{\varphi}_n\|_{L^2}$$

for all $\mathbb{X} \supset \operatorname{supp}(w_\alpha)$. In the case of the Radon transform, we can compute the left-hand side explicitly for the finitely many $j \in J_n$ but we wish to use the right-hand side in a structured way to avoid computing the infinitely many $j \notin J_n$. To do this, we will take a geometrical perspective on the construction of wavelets to view them in a tree format.

**Tree structure of wavelets** Finite elements are constructed with a mesh which provided a useful tool for adaptive refinement in Section 4.6.3. For wavelets, we will associate a tree with every discretisation and the leaves of the tree form a mesh. This perspective comes from the multi-resolution interpretation of wavelets. We will explain the approach for 1D in detail and then comment on how to extend this picture to higher dimensions. We start with a space $\widetilde{\mathbb{U}}^0$ and a normalised mother wavelet $\psi\colon [0,1] \to \mathbb{R}$ then inductively form $\widetilde{\mathbb{U}}^k$ by

$$\widetilde{\mathbb{U}}^k = \widetilde{\mathbb{U}}^{k-1} \bigcup \left\{w_{j,k}(x) = \sqrt{2}^k\psi(2^kx - j) \text{ s.t. } j = 0, \dots 2^k - 1\right\}.$$

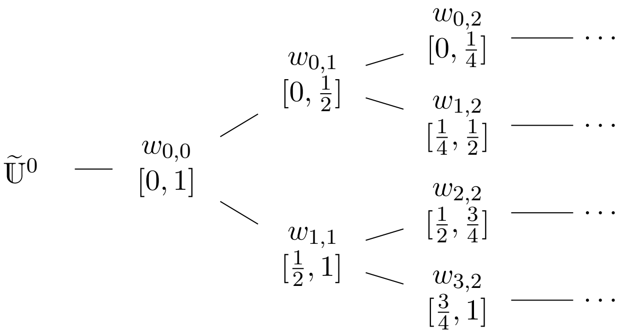


**Figure 4.5** Tree representation of 1D wavelets $w_{j,k}$ are arranged in a tree structure with their support underneath.

If we keep track of the support $w_{j,k}$ then we see a tree structure emerging, as shown in Figure 4.5. Each time a node splits, the support is partitioned exactly between its two children. If we truncate this tree to $J_n$ such that every node either has zero or two children, then the leaves of this tree form a partition of unity. For example

$$\begin{aligned} J_n = \{(0,0),(0,1),(1,1),(2,2),(3,2)\} \implies \operatorname{leaf}(J_n) &= \{(0,1),(2,2),(3,2)\}, \\ [0,1] &= [0,\tfrac12]\cup[\tfrac12,\tfrac34]\cup[\tfrac34,1]. \end{aligned}$$

In higher dimensions, the only two things which change are the number of children ($2^d$ for non-leaves) and at each node you store the coefficients of $2^d-1$ wavelets. The support on each node is still a disjoint partition of unity consisting of regular cubes of side length $2^{-k}$ at level $k$. The only change in our own implementation is to translate the support to $[-\frac12,\frac12]^2$. We briefly remark that the tree structuring of wavelets is not novel and appears more frequently in the Bayesian inverse problems literature, for example in Castillo and Rockova (2019).

**Continuous gradient estimate** In Section 4.7.1 we used the continuous gap as a measure for convergence, for wavelets we will use the continuous gradient. With the tree structure we can easily adapt the results of Section 4.6.3 to estimate gradients (or function gaps). In particular,

$$\begin{aligned} |||\partial \operatorname{E}(u_n)|||_* &= \max\left(|||\partial_n \operatorname{E}(u_n)|||_*, \max_{j\notin J_n} |\langle w_j,\ \mathcal{A}^*\boldsymbol{\varphi}_n\rangle| - \mu\right) && (4.9)\\ &\le \max\left(|||\partial_n \operatorname{E}(u_n)|||_*, \max_{j\in\operatorname{leaf}(J_n)} \left\|\mathbb{1}_{\operatorname{supp}(w_j)}\mathcal{A}^*\boldsymbol{\varphi}_n\right\|_{L^2} - \mu\right). && (4.10)\end{aligned}$$

It is interesting to note that

$$\sum_{j\in\operatorname{leaf}(J_n)} \left\|\mathbb{1}_{\operatorname{supp}(w_j)}\mathcal{A}^*\boldsymbol{\varphi}^*\right\|_{L^2}^2 = \|\mathcal{A}^*\boldsymbol{\varphi}^*\|_{L^2}^2,$$

therefore the task of refinement is somehow to partition the domain of $\mathcal{A}^*\boldsymbol{\varphi}^*$ such that no single component has more than $\mu$ magnitude. Also, by strong convexity of the dual problem,

$$\left|\gamma_0\left\|\mathbb{1}_{\operatorname{supp}(w_j)}\mathcal{A}^*\boldsymbol{\varphi}_n\right\|_{L^2}-\left\|\mathbb{1}_{\operatorname{supp}(w_j)}\mathcal{A}^*\boldsymbol{\varphi}^*\right\|_{L^2}\right|\leq\sqrt{|\operatorname{supp}(w_j)|}\sqrt{2(\mathrm{E}(u_n)+\mathrm{E}^*(\gamma_0\boldsymbol{\varphi}_n))},$$

therefore the exact discretisation is reached in finite time. After this point, the discretised problem is equal to the continuous problem, and Algorithm 4.1 will behave as in the classical setting.

**Numerical results** We consider two phantoms where $u^\dagger$ is either a binary disc or the Shepp-Logan phantom. No noise is added to the Shepp-Logan data but 5 % Gaussian white noise is added to the disc data. This is visualised in Figure 4.6. All optimisations shown will be spatially adaptive using Haar wavelets and initialised with four degrees of freedom (denoted $\mathbb{U}^1$ in the notation of Figure 4.5). The gradient metric shown throughout is the $\ell^2$ norm. Motivated by (4.10), the spatial adaptivity is chosen to refine nodes $j\in\operatorname{leaf}(J_n)$ such that

$$\left\|\mathbb{1}_{\operatorname{supp}(w_j)}\mathcal{A}^*\boldsymbol{\varphi}_n\right\|_{L^2}-\mu\leq 10|||\partial_n\mathrm{E}(u_n)|||_*,$$

i.e. so that the continuous gradient is less than 10 times the discrete gradient.

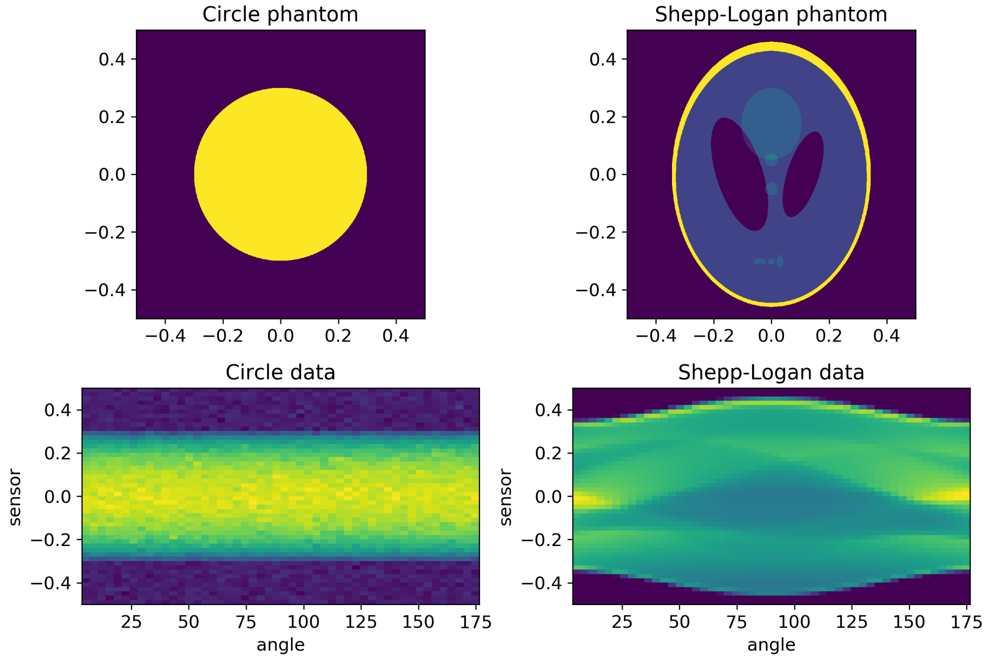


**Figure 4.6** Phantoms and data used for wavelet-sparse tomography optimisation. The Shepp-Logan data is exact but the data for the disc-phantom has 5 % Gaussian white noise. Without noise the data would be uniform with respect to the angle.

The first numerical results shown in Figure 4.7 compare the same adaptive algorithms as shown in Figure 4.2. In these examples we see that the greedy FISTA, restarting, and the $a = 20$ algorithms achieve almost linear convergence while $a = 2$ and the classical FB are significantly slower. The maximum number of wavelet coefficients used was 312,220 and 44,644 for the circle and Shepp-Logan phantoms respectively.

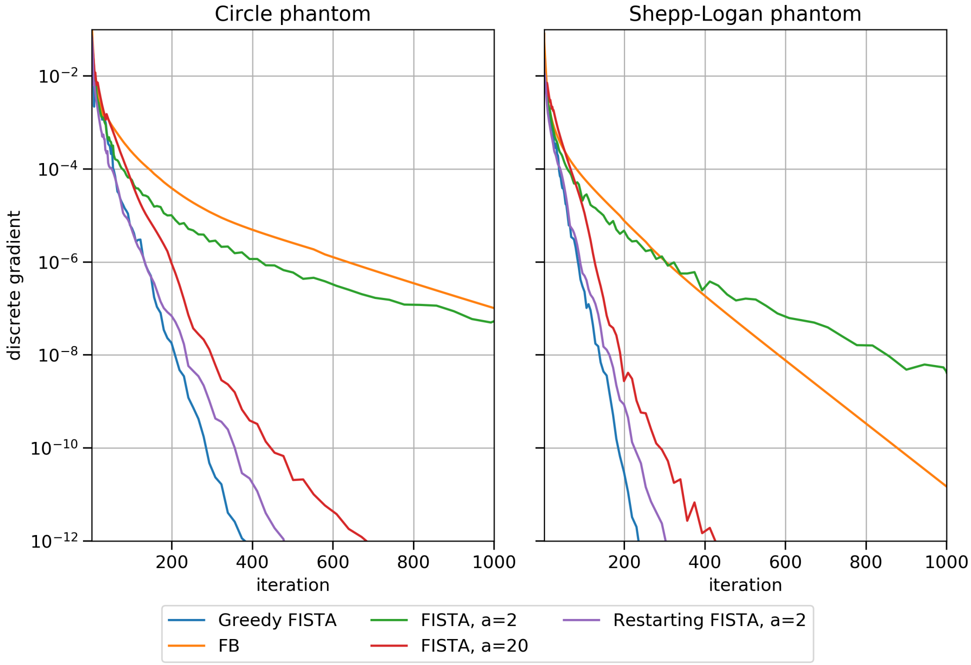


**Figure 4.7** Discrete convergence of different implementations of Algorithm 4.1 with an unlimited number of pixels.

As before, we focus on the $a = 20$ algorithm to which Lemma 4.4.11 applies, and greedy FISTA which we see achieves slightly faster convergence in Figure 4.8. Looking at the discrete and continuous gradient norms, we see that they are initially distinct then merge after around 50 iterations. From this point onwards, the continuous and discrete problems are equivalent and the iterations are equivalent to classical FISTA.

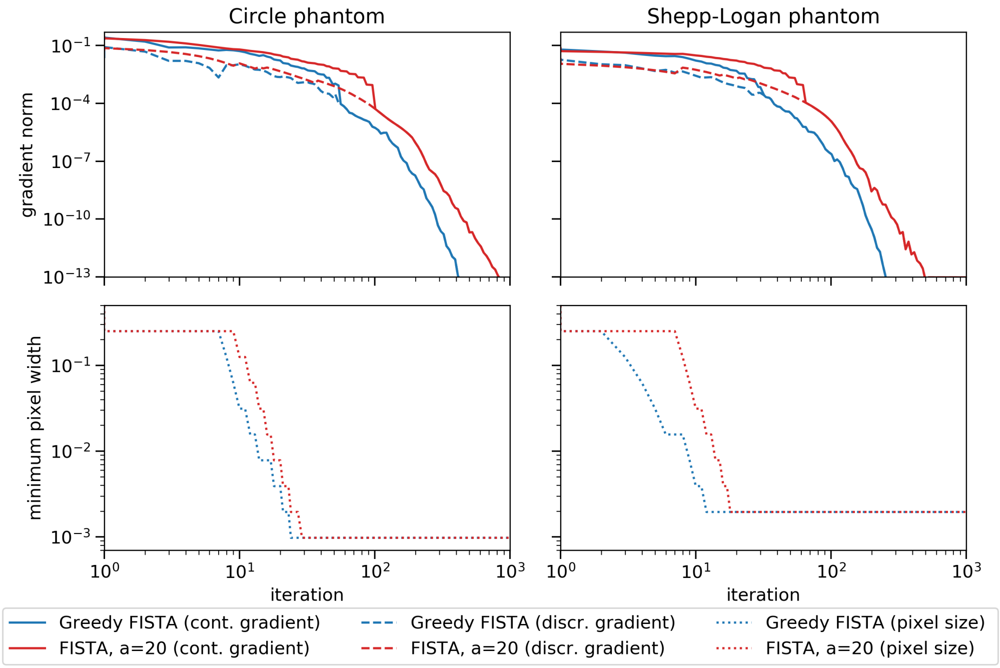


**Figure 4.8** Discrete/continuous convergence of adaptive FISTA algorithms on the wavelet Lasso optimisation.

### 4.7.3 2D continuous Lasso

Our final application is a super-resolution/de-blurring inverse problem from biological microscopy. The resolution of visible light microscopy is fundamentally limited by the wavelength of visible light. Classically, this limit has held around 250 nm, however, in recent years new techniques have emerged improving resolution to around 30 nm to 50 nm (Schermelleh et al., 2019). A big component of this shift is a growing reliance on more powerful data processing techniques. *Stochastic Optical Reconstruction Microscopy* (STORM) is an example of *Single Molecule Localisation Microscopy* (SMLM) where a large number of coarse blurred images are recorded, then re-combined to form a single sparse super-resolved image. In the context of STORM, each recorded image is modelled as a sparse signal convolved with a point-spread function, then corrupted with noise. The Lasso formulation has previously been shown to be effective in the context of STORM (Huang et al., 2017; Denoyelle et al., 2019).

In this example we use a simulated dataset provided as part of the 2016 SMLM challenge[1] for benchmarking software in this application. It is common to model the point-spread function as a Gaussian, in this example the corresponding Lasso formulation is

$$(\mathcal{A}u)_i = (2\pi\sigma^2)^{-1}\int_{[0,6.4]^2} \exp\left(-\frac{1}{2\sigma^2}\left|\boldsymbol{x} - \Delta\begin{pmatrix} i_1+\frac{1}{2} & i_2+\frac{1}{2}\end{pmatrix}^\top\right|^2\right) u(\boldsymbol{x})$$

[1] http://bigwww.epfl.ch/smlm/challenge2016/datasets/MT4.N2.HD/Data/data.html

where $\sigma = 0.2\,\mu\mathrm{m}$, $\Delta = 0.1\,\mu\mathrm{m}$, and $i_1, i_2 \in [64]$ where $\mathbb{U} = \mathcal{M}([0\,\mu\mathrm{m}, 6.4\,\mu\mathrm{m}]^2)$. 3020 frames are provided, examples of which are shown in Figure 4.9. To process this dataset, image intensities were normalised to $[0,1]$ then a constant was subtracted to approximate 0-mean noise. The greedy FISTA algorithm was used for optimisation with $\mu = 0.15$, $10^3$ iterations, and a maximum of $10^5$ pixels per image.

Finally, all the reconstructions were summed and the result shown in Figure 4.10. The average pixel width after optimisation was approximately 2.4 nm, a factor of 40 super-resolution. If this resolution had been implemented with a uniform discretisation then it would have required $70 \times 10^5$ pixels, nearly a factor of 100 greater than achieved with the adaptive discretisation. Lasso is compared with ThunderSTORM (Ovesnỳ et al., 2014), a popular ImageJ plugin (Schindelin et al., 2012) which finds the location of signal using Fourier filtering. The performance of ThunderSTORM was rated very highly in the initial competition by Sage et al. (2015). Both methods compared here demonstrate the key structures of the reconstruction, however, both are sensitive to tuning parameters. In this examples, Lasso has possibly recovered too little signal and ThunderSTORM contains spurious signal.

Figure 4.11 shows the convergence behaviour in this example. The estimates given by (4.6) and (4.7) in dimension $d = 2$ predict respectively that the adaptive energy will decay at a rate of $\mathrm{E}_0(u_n) \lesssim n^{-2/3}$ so long as the pixel size also decreases at a rate of $h \sim n^{-2/3}$. This is consistent with the resolution scaling (middle panel) but the energy (left panel in pink) is observed to converge a little faster than predicted.

In this example we also implement the suggestion of Section 4.6.4 to remove pixels from the iteration once we can guarantee they are outside of the support. (4.8) provides a threshold to identify the support of the discrete/continuous minimiser and the value of this is plotted in the first panel of Figure 4.11, in particular the normalised value $1 - \frac{\text{threshold}}{\mu}$ which converges to 0 for large $n$. Any pixel $\omega_i^n$ satisfying

$$\gamma_0 \left\| \Pi_n \mathcal{A}^* \boldsymbol{\varphi}_n \right\|_{L^\infty(\omega_i^n)} \leq \text{threshold}$$

guarantees that $\omega_i^n \cap \operatorname{supp}(u^*) = \emptyset$. Once this threshold becomes greater than 0 (plotted value less than 1), we can start reducing the number of pixels instead of just continual refinement. We can see this in the right-hand panel of Figure 4.11, after around 30 iterations the total number pixels starts to reduce and stabilise at approximately $6 \times 10^3$ pixels per frame, well below the upper limit of $10^5$.

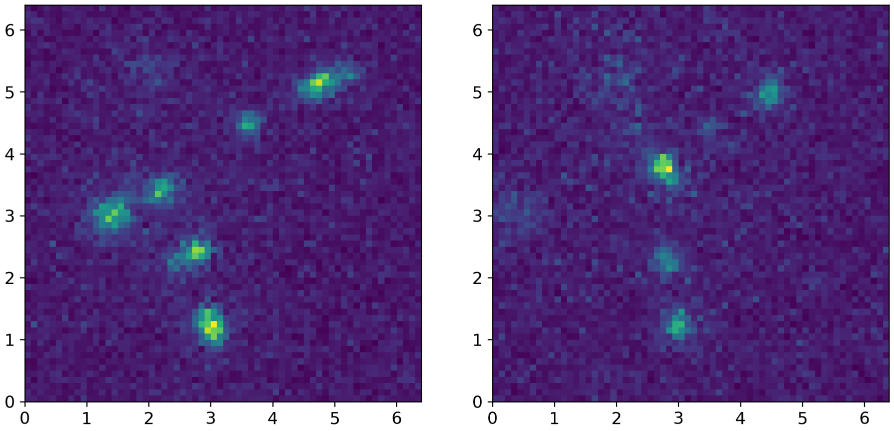


**Figure 4.9** Example of images from STORM dataset.

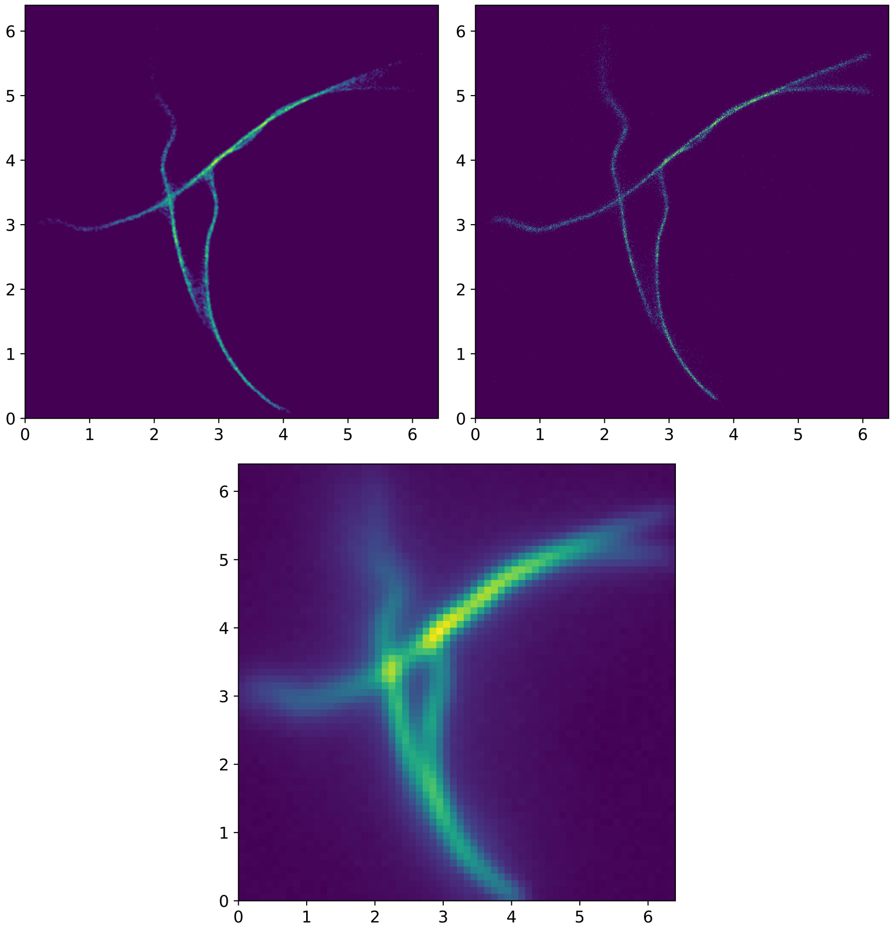


**Figure 4.10** Processed results of the STORM dataset. Top left: Lasso optimisation with Algorithm 4.1. Top right: Comparison with ThunderSTORM plugin. Bottom: Average data, no super-resolution or de-blurring.

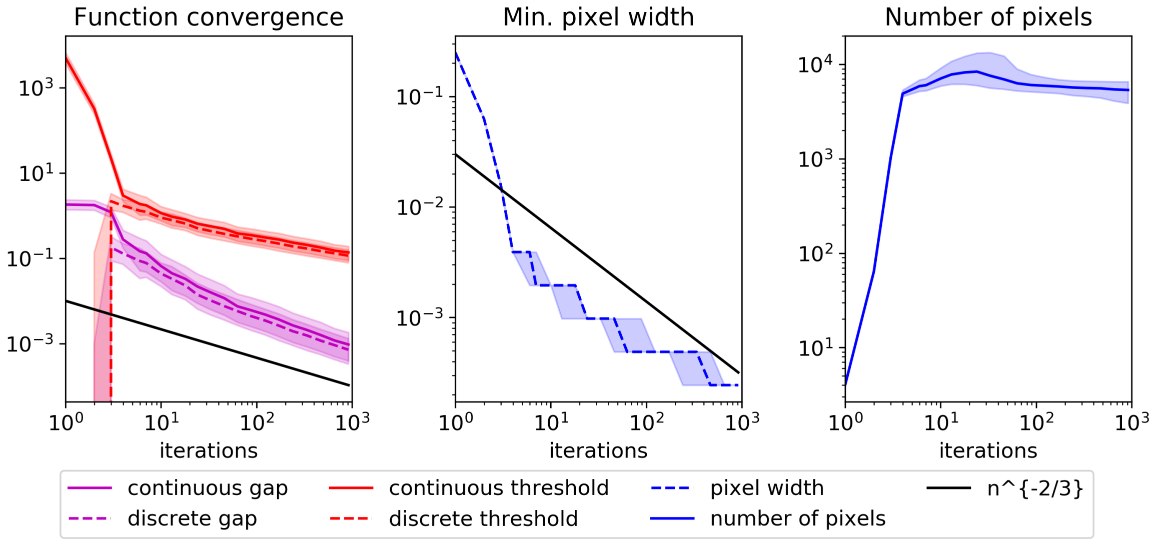


**Figure 4.11** Convergence of adaptive FISTA for STORM dataset. Lines indicate the median value over 3020 STORM frames. Shaded regions indicate the 25 % to 75 % inter-quartile range.

## 4.8 Conclusions and outlook

In this work we have proposed a new adaptive variant of FISTA and provided convergence analysis. This algorithm allows FISTA to be applied outside of the classical Hilbert space setting, still with a guaranteed rate of convergence. We have presented several numerical examples where convergence with the refining discretisation is at least as fast as a uniform discretisation, although more efficient with regards to both memory and computation time.

In 1D we see good agreement with the theoretical rate. This rate also seems to be a good predictor for all variants of FISTA tested, although this is yet to be proven. Surprisingly, even the classical methods with a fixed discretisation initially seem limited to the slower adaptive rate for small $n$.

The results in 2D are similar, all tested FISTA methods converge at least at the guaranteed rate. The wavelet example was most impressive, achieving nearly linear convergence in energy. This is similar to the behaviour for classical FISTA although it is also yet to be formally proven.

An interesting observation over all of the adaptive Lasso examples is that the classical oscillatory behaviour of FISTA has not occured. With the monotone gaps plotted, oscillatory convergence should correspond to a piecewise constant descending gap. Either this behaviour only emerges for larger $n$, or the adaptivity mimics the restarting behaviour typically used to avoid oscillation. The perturbation provided by the refinement is sufficient to stop FISTA overshooting the minimiser and maintain a predictable rate of convergence to the minimiser. This may also explain why the standard restarting FISTA showed little improvement in these examples.

Moving forward, it would be interesting to see how far the analysis extends to other optimisation algorithms. To cover the numerical results, this would necessitate repeating the presented argument for Forward-Backward splitting and for the modifications of FISTA suggested by Liang and Schönlieb (2018). We feel that the extension to the primal-dual algorithm proposed by Chambolle and Pock (2011) should also be possible where there is a classical bound of the form $\min_{n\leq N} \mathrm{E}(u_n) + \mathrm{E}^*(\boldsymbol{\varphi}_n) \lesssim \frac{\|u_0-u^*\|^2+\|\varphi_0-\boldsymbol{\varphi}^*\|^2}{N}$. This has the same basic ingredients as FISTA although there is now the extra dual variable to account for.

Another interesting outlook is to relax the requirement for f to be smooth on the whole of $\mathbb{U}\cap\mathbb{H}$. An example where this is not the case is the dual problem to total variation denoising where we have

$$\mathrm{E}(u) = \|\operatorname{div} u - \boldsymbol{\eta}\|^2 + \chi(\|u\|_\infty \leq \mu).$$

The divergence operator is not bounded in $L^2$ but is on each subspace $\mathbb{U}^n$. It is not clear if it is possible to incorporate this into FISTA without restarting each time the step-size changes.

# Chapter 5

# Total Variation Discretisation

Total variation is a very common regularisation functional in inverse problems. One reason for this is its analytical properties, reconstructions are guaranteed to be piecewise constant with 'nice' level sets. On the other hand, to approximate these reconstructions one must choose a discretisation, which in turn dictates which of these properties are realised numerically.

The most common discretisation is with finite differences, although this also has some of the weakest approximation properties to the continuous problem (Bartels, 2012; Condat, 2017). Many finite difference-like alternatives have been proposed which may not have strong analytical links to the original TV functional, but ensure that discrete reconstructions share the same heuristic properties as the continuum (Chambolle et al., 2009; Condat, 2017). Another approach is to use a finite element discretisation, such as piecewise linear, which have been shown to have much better analytical guarantees than the standard finite differences (Bartels, 2012, 2015; Chambolle and Pock, 2020).

The key feature that we will focus on in is the rate of approximation for different finite element discretisations of total variation. Recent work by Chambolle and Pock (2020) introduced a discretisation which achieves a new and faster rate when certain technical assumptions are met. This is accompanied with a conjecture that the necessary assumptions are always satisfied in practice. In this chapter we demonstrate that the conjecture is not true in general, and then consider the construction of a finite element which could achieve the fast rate for any minimiser.

## 5.1 Background

The prototypical inverse problem that is considered when investigating properties of total variation regularisation is total variational denoising. The functional we seek to minimise is

$$\mathrm{E}(u) = \tfrac{1}{2}\left\|u-\eta\right\|_2^2 + \mu\,\mathrm{TV}(u), \qquad u^* := \operatorname*{argmin}_{u\in\mathbb{BV}(\Omega)} \mathrm{E}(u) \tag{5.1}$$

for some $\eta \in L^\infty(\Omega)$. There are two key reasons for choosing the denoising problem. Firstly, the problem is strongly convex so

$$\tfrac{1}{2}\left\|u-u^*\right\|_2^2 \leq \mathrm{E}(u)-\mathrm{E}(u^*).$$

i.e. bounds on the energy also give $L^2$ bounds on the reconstruction approximation. The other reason is that many general optimisation algorithms, such as Algorithms 1.1 and 1.4, only require computation of the proximal maps. For example, Algorithm 1.1 applied to (1.2) with TV regularisation would give

$$u_{n+1} = \operatorname*{argmin}_u \tfrac{1}{2}\left\|u-u_n+\|\mathcal{A}\|^{-2}\mathcal{A}^*(\mathcal{A}u_n-\eta)\right\|_2^2 + \mu\,\|\mathcal{A}\|^{-2}\,\mathrm{TV}(u).$$

Each iteration of the algorithm can be understood by understanding (5.1). Also, the minimiser is a fixed point of this algorithm so $u^*$ must coincide with the solution of (5.1) for some modified data.

Moving on to discretisation, we are looking for minimisation problems of the form

$$u_h^* \coloneqq \operatorname*{argmin}_{u\in\mathbb{U}_h} \mathrm{E}_h, \qquad \mathrm{E}(u_h^*) \approx \mathrm{E}(u^*). \tag{5.2}$$

where $\mathbb{U}_h$ is a discretisation with mesh size $h$. If $\mathrm{E}_h = \mathrm{E}$, then the discretisation is called a *conforming* finite element method. There are three key results to highlight:

- If $\mathbb{U}_h$ is a space of piecewise constant finite elements, then the conforming discretisation is only Gamma-convergent (Bartels, 2012, 2015). In particular, let $\eta(x,y) = \mathbb{1}_{x+y\leq 0}$ and $\mathbb{U}_h$ the set of piecewise constant functions on a uniform mesh of squares of width $h$. Then, as $h\to 0$, $u_h^*$ converges to $u^*$ in $L^1$ (but not in $\mathbb{BV}$) and $\mathrm{E}(u_h^*)$ *does not* converge to $\mathrm{E}(u^*)$.

- If $\mathbb{U}_h$ is a space of continuous piecewise linear finite elements, then the conforming discretisation converges with

  $$\mathrm{E}(u_h^*)-\mathrm{E}(u^*) \lesssim \sqrt{h}.$$

  i.e. piecewise linear elements converge with $\|u_h^*-u^*\|_2 \lesssim h^{\frac{1}{4}}$ (Bartels, 2015).

- If $\mathbb{U}_h$ is the space of *Crouzeix-Raviart* piecewise linear finite elements and $\mathrm{E}_h$ is chosen to be non-conforming, then

  $$\mathrm{E}(u_h^*)-\mathrm{E}(u^*) \lesssim h$$

  whenever some technical assumption is satisfied (Chambolle and Pock, 2020).

This technical assumption will be stated formally in Section 5.2 where we also give an example where it does not hold, and indeed the fast rate is not achieved. In Section 5.3 we begin to construct a finite element which might achieve the faster rate for any $\eta \in L^\infty$.

In this work we focus on the case where $\Omega$ is a bounded subset of $\mathbb{R}^2$.

## 5.2 Counter example

An important tool in analysing (5.1) is the dual form of the optimisation problem. This is derived more thoroughly by Chambolle and Pock (2020); Bartels (2020), however, we shall simply state that the dual problem is

$$\mathrm{E}^*(\varphi) = \tfrac{1}{2}\|\mathrm{div}\,\varphi\|_2^2 + \langle \eta,\ \mathrm{div}\,\varphi\rangle + \chi(\|\varphi\|_{\infty,2} \leq \mu) + \chi(\varphi\boldsymbol{\cdot}\nu|_{\partial\Omega} = 0) \tag{5.3}$$

where $u^* = \mathrm{div}\,\varphi^* + \eta$ holds in $L^2$, $\nabla u^*\boldsymbol{\cdot}\varphi^* = |\nabla u^*|$ holds in the sense of distribution, and $\nu\colon \partial\Omega \to \mathbb{R}^2$ is the oriented boundary normal.

The convergence result first shown by (Chambolle and Pock, 2020) and generalised slightly by (Bartels, 2020) is stated as follows.

**Theorem 5.2.1** ((Chambolle and Pock, 2020, Section 5.1.1), (Bartels, 2020, Proposition 4.2))**.** *Let $\mathbb{U}_h$ be a space of piecewise linear Crouzeix-Raviart finite elements with a mesh of uniform pixel diameter $h$. If $\eta \in L^\infty$, $u^* = \operatorname{argmin} \mathrm{E}$ and*

$$\textit{there exists} \quad \varphi^* \in \operatorname{argmin} \mathrm{E}^* \quad \textit{such that} \quad \varphi^* \textit{ is Lipschitz,} \tag{5.4}$$

*then*

$$\mathrm{E}(u_h^*) - \mathrm{E}(u^*) \lesssim h.$$

Chambolle and Pock (2020) comment that the smoothness of $\varphi$ can be related to the smoothness of the level sets of $u^*$, which are in turn inherited from the level sets of $\eta$. It is conjectured that some smoothness assumption on the level sets of $\eta$ would guarantee the existence of Lipschitz $\varphi^*$. Here we give an example with smooth level sets where the dual function cannot be Lipschitz. The example is sketched in Figure 5.1. In essence, the level sets are smooth but their union forms a cusp which will be a point where $\varphi^*$ is not Lipschitz.

It now remains to provide a formal argument which first shows that Figure 5.1 is accurate, then confirms that the consequence is a point of non-Lipschitz continuity.

**Theorem 5.2.2.** *If*

$$\eta(x,y) = \begin{cases} +1 & x^2 + (1-y)^2 < 1 \\ -1 & x^2 + (-1-y)^2 < 1 \\ 0 & \textit{else} \end{cases} \qquad \textit{for all } x, y \in [-2,2],$$

*then*

$$u^* = \max(0, 1-2\mu)\eta.$$

*Proof.* We first claim that, because $\eta$ is an odd function in $y$, $u^*$ must also be odd in $y$. This follows by strong convexity:

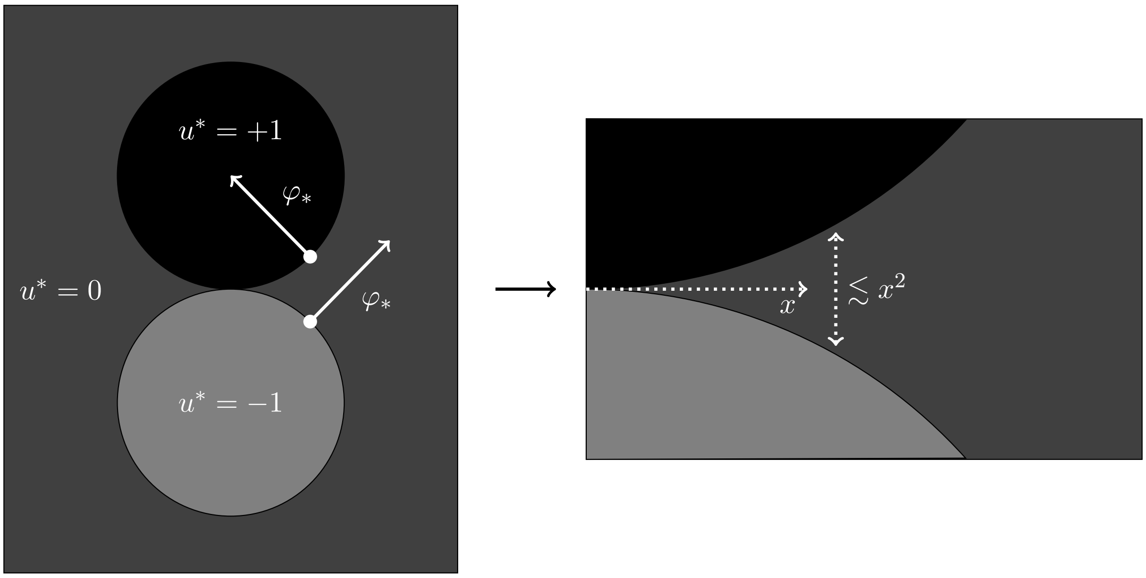


**Figure 5.1** Example of a reconstruction where the dual function is not Lipschitz-continuous. The left-hand plot shows $\eta$ (also a scalar multiple of $u$). On the level sets of $u^*$, $\varphi^*$ must be a unit normal in the direction of the jump. The right-hand plot highlights that the gap between level sets vanishes faster than the distance to meeting. This forces $\varphi^*$ to be non-Lipschitz.

$$\begin{aligned}\mathrm{E}\left(\tfrac{u^*(x,y)-u^*(x,-y)}{2}\right) &= \frac{1}{2}\left\|\tfrac{u^*(x,y)-u^*(x,-y)}{2} - \tfrac{\eta(x,y)-\eta(x,-y)}{2}\right\|_2^2 + \mu\,\mathrm{TV}\left(\tfrac{u^*(x,y)-u^*(x,-y)}{2}\right)\\ &\leq \frac{\frac{1}{2}\left\|u^*-\eta\right\|_2^2 + \frac{1}{2}\left\|u^*-\eta\right\|_2^2}{2} - \frac{\left\|u^*(x,y)+u^*(x,-y)\right\|_2^2}{8} + \frac{\mu}{2}\left(\mathrm{TV}(u^*)+\mathrm{TV}(u^*)\right)\\ &= \mathrm{E}(u^*) - \frac{1}{8}\left\|u^*(x,y)+u^*(x,-y)\right\|_2^2.\end{aligned}$$

Therefore, every minimiser $u^*$ must also satisfy $u^*(x,y) + u^*(x,-y) = 0$.

If $u^*$ is odd in $y$, then $u^*(x,0) = 0$ in the trace sense and we can focus on the half-plane $y > 0$. The half-plane problem is very standard and we can use the following standard result (e.g. Chambolle et al., 2016, Equation 13)

$$u_+^* = \operatorname*{argmin}_{u\in\mathbb{BV}(\mathbb{R}^2)} \tfrac{1}{2}\left\|u - \eta\mathbb{1}_{y\geq 0}\right\|_2^2 + \mu\,\mathrm{TV}(u) \qquad \iff \qquad u_+^* = \max(0, 1-2\mu)\eta\mathbb{1}_{y\geq 0}.$$

Note that $u_+^*(x,0)=0$ in the $L^1(\mathbb{R})$ sense, therefore

$$\begin{aligned}
\mathrm{E}(u^*) &= \min_{u(x,0)=0} \tfrac{1}{2}\left\|u-\eta\right\|_2^2 + \mu\int |\nabla u| \\
&\geq \min_{u(x,0)=0} \tfrac{1}{2}\left\|u-\eta\right\|_2^2 + \mu\int_{y>0} |\nabla u| + \mu\int_{y<0} |\nabla u| \\
&= 2\min_{u(x,0)=0} \tfrac{1}{2}\left\|u-\eta\right\|_{L^2(y>0)}^2 + \mu\int_{y>0} |\nabla u| \\
&= 2\left(\tfrac{1}{2}\left\|u_+^*-\eta\mathbb{1}_{y>0}\right\|_2^2 + \mu\int |\nabla u_+^*|\right) \\
&= \mathrm{E}(u_+^*(x,y)-u_+^*(x,-y)) \\
&\geq \min_{u\in\mathbb{U}} \mathrm{E}(u) = \mathrm{E}(u^*).
\end{aligned}$$

The first inequality is because we have reduced the energy by ignoring jumps on the $y=0$ axis. This separates the optimisation into two problems to which we already know $u_+^*$ is feasible and minimises each sub-problem. We can then combine the problems again with the fact that $\mathrm{E}(u^*)$ is minimal. This shows that $\mathrm{E}(u^*)=\mathrm{E}(u_+^*(x,y)-u_+^*(x,-y))$ and so the unique minimiser is

$$u^* = u_+^*(x,y)-u_+^*(x,-y) = \max(0,1-2\mu)\left[\eta\mathbb{1}_{y\geq 0}-\eta(x,-y)\mathbb{1}_{y\leq 0}\right] = \max(0,1-2\mu)\eta.$$

□

Now we must show that every dual solution $\varphi^*$ is non-Lipschitz at some point. The non-uniqueness of the dual solution is another potential challenge, although we bypass this with the criticality condition $\nabla u^*\cdot\varphi^*=|\nabla u^*|$. This has no influence outside of the support of $\nabla u^*$ but the chosen jump-set is still sufficient.

**Lemma 5.2.3.** *If $u^*=\alpha\eta$ for some $\alpha>0$, then $\varphi^*$ is at most $\frac{1}{2}$-Hölder continuous. In particular, it must be not globally Lipschitz.*

*Proof.* Recall that the general definition of Hölder continuity is

$$\varphi\in C^\theta \qquad\iff\qquad \limsup_{\boldsymbol{r},\boldsymbol{r}'} \frac{|\varphi(\boldsymbol{r})-\varphi(\boldsymbol{r}')|}{|\boldsymbol{r}-\boldsymbol{r}'|^\theta} < \infty.$$

We shall lower-bound this limit in the neighbourhood of 0, considering the level sets drawn in the right-hand panel of Figure 5.1.

The condition $\nabla u^*\cdot\varphi^*=|\nabla u^*|$ (Chambolle et al., 2016, Proposition 14) fixes the values of $\varphi^*$ on the jump-set:

$$\varphi^*(x,y)=\begin{cases} (0-x,1-y) & (0-x)^2+(1-y)^2=1 \\ -(0-x,-1-y) & (0-x)^2+(-1-y)^2=1 \end{cases}.$$

The sign of the unit normal indicates the direction in which $u^*$ is increases on the jump-set. The rest is direct computation. Let

$$y(x) = 1 - \sqrt{1-x^2},$$

then for all $\theta \leq 1$

$$\begin{aligned}\frac{|\varphi(x,y(x)) - \varphi(x,-y(x))|^2}{|(x,y(x)) - (x,-y(x))|^{2\theta}} &= \frac{(2x)^2+(2y)^2}{(4y^2)^\theta} \\ &= 4^{1-\theta}\frac{x^2}{(1-\sqrt{1-x^2})^{2\theta}} + 4^{1-\theta}y^{2-2\theta} \\ &\sim 4^{1-\theta}\frac{x^2}{(1-1+\frac{1}{2}x^2)^{2\theta}} && x \ll y \\ &= 4^{1-\theta}x^{2-4\theta}.\end{aligned}$$

This is unbounded for all $\theta > \frac{1}{2}$. □

This argument demonstrates that $u^*$ can be only $\frac{1}{2}$-Hölder continuous at points of level set intersection. In the continuous setting, it seems natural that there should exist some smoothness assumption on $\eta$ which avoids this scenario. However, Boyer et al. (2019) show that all TV minimisers with finite data must be built up of a finite number of indicators on 'simple' sets. This suggests that the poor continuity property should be common in practice.

We have now shown analytically that the proof for a faster rate does not always hold but it is also of interest to see whether or not the rate is still achieved numerically. To test this, we simulated data with corresponding exact reconstruction seen in Figure 5.2. This is slightly different to the simple two-disc phantom considered above but a similar proof can be repeated.

Figure 5.3 is an extension to that shown by Bartels (2020). For this reason we plot $h$ against $\|u_h^* - u^*\|_2^2$ as a proxy of the function gap convergence. The original $\eta$ in Bartels (2020) was a simple single disc indicator function. The important comparison is the fact that the slopes on the left-hand plot are all twice as steep as on the right-hand. The purple line indicates the theoretical rate although the observed rate appears to be a little slower. The key take-home from this figure is that the observed convergence of the solution which does not satisfy (5.4) is twice as slow as the rate for the example which does satisfy the condition.

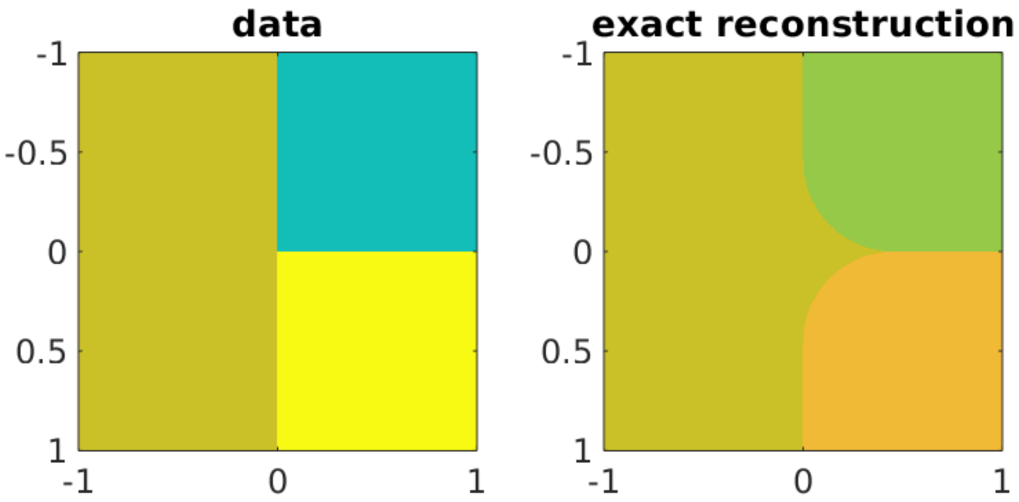


**Figure 5.2** Original data (left) with exact reconstruction (right). The right-angled meeting point is smoothed to circular arcs and the contrast becomes reduced.

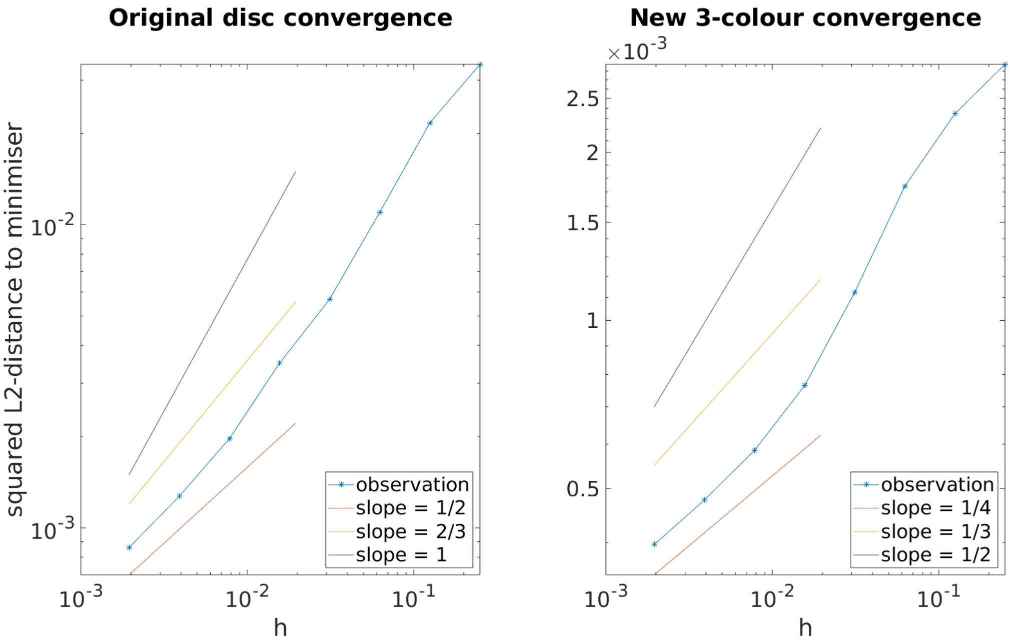


**Figure 5.3** Convergence of $L^2$ distance with respect to $h$. If (5.4) is satisfied, then we expect the convergence to be of rate $h$, otherwise it should be of order $\sqrt{h}$. Note that the $y$-axis scaling is different in each plot, slopes with equal colour are actually a factor of two shallower in the right-hand plot.

## 5.3 TV finite element

### 5.3.1 Problem formulation

In this section we consider construction of a new type of finite element which would perform well for the examples motivated by Theorem 5.2.2 and Lemma 5.2.3. Our aim is to produce a discretisation of (5.1) such that:

- $u_h^*$ is piece-wise constant,
- $\mathrm{E}(u_h^*) - \mathrm{E}(u^*) \lesssim h$,
- and $u_h^*$ is still the minimiser of a *convex* optimisation problem.

Current bases which achieve order $h$ errors are piecewise linear but fail to maintain this rate when $u^*$ is piece-wise constant and not sufficiently smooth (Chambolle and Pock, 2020; Bartels, 2020). On the other hand, Bartels (2012, 2015) shows that naive piece-wise constant discretisation is also not powerful enough to achieve order $h$ convergence, even when $u^*$ is smooth. We hope to avoid this limitation by allowing the piecewise constant mesh of $u_h^*$ to adapt to $u^*$; the challenge is to preserve convexity.

It is the hope that such a basis will achieve a faster rate of convergence, overcoming the difficulties in Section 5.2. Again, this is still preliminary work and so there are not yet any proofs to support this claim, but we will first provide a motivation for our approach.

### 5.3.2 Graphical rate estimation

The aim of this section is to produce a finite element which achieves an error rate of order $h$. In this subsection, we outline the key obstacles which need to be overcome to achieve this rate. This total error is the summation of errors over each pixel and so we can begin to categorise each pixel by difficulty, then estimate the necessary rate in each case. This subsection is not at all rigorous but intended to highlight the key challenges a proof would need to address.

To analyse the error contributed by each pixel, we introduce a normalised metric:

$$\varepsilon_i := \underbrace{\|u_h^* - u^*\|_{L^1(\omega_i)}}_{\text{data term}} + \underbrace{\mu\left[\int_{\omega_i} |\nabla u_h^*| - |\nabla u^*|\right]}_{\text{interior TV}} + \underbrace{\tfrac{\mu}{2}\left\|\mathrm{Tr}_{\partial\omega_i^+} u_h^* - \mathrm{Tr}_{\partial\omega_i^-} u_h^*\right\|_{L^1(\partial\omega_i)}}_{\text{boundary TV}}. \tag{5.5}$$

The final term measures jumps of intensity which coincide with the pixel boundaries. This term is almost surely zero for $u^*$ but the discretisation may not be, for example a standard piece-wise constant discretisation can only jump on the boundaries. This error is chosen because it is a

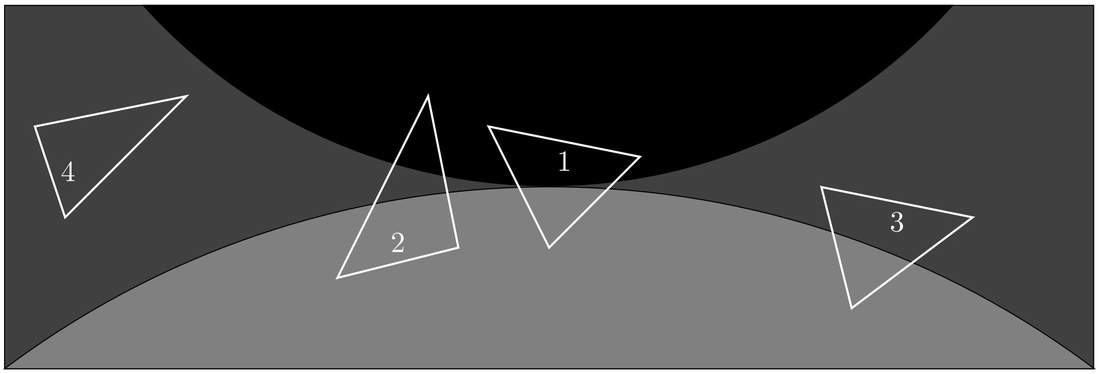


**Figure 5.4** A good convergence rate requires good estimation of $u^*$ on four types of domain characterised by regularity of the dual variable $\varphi^*$. The Lipschitz constant is of order $\infty$, $\frac{1}{h}$, 1 in pixels 1 to 3 respectively. Pixel 4 is just a constant value.

simple decomposition over pixels and produces the upper bound:

$$\begin{aligned}\mathrm{E}(u_h^*) - \mathrm{E}(u^*) &\leq (2\left\|\eta\right\|_\infty + \left\|u_h^*\right\|_\infty + \left\|u_h^*\right\|_\infty)\left\|u_h^* - u^*\right\|_1 + \mu(\mathrm{TV}(u_h^*) - \mathrm{TV}(u^*)) \\ &\lesssim \left\|u_h^* - u^*\right\|_1 + \mu(\mathrm{TV}(u_h^*) - \mathrm{TV}(u^*)) = \sum_i \varepsilon_i.\end{aligned}$$

The task is now to find a discretisation which bounds $\varepsilon_i$ as small as possible. In the case of naive piece-wise constant discretisation, we get $\varepsilon_i \lesssim h\left\|u^*\right\|_\infty$. This error comes from the boundary TV term ($O(h)$) rather than the data or interior TV term which scale with $|\omega_i| = O(h^2)$.

To derive more tight bounds, it is natural that more complicated pixels are harder to approximate (see Figure 5.4). The natural measure for this complexity is $\mathrm{TV}_{\omega_i}(u^*) \coloneqq \int_{\omega_i} |\nabla u^*|$: if the pixel is constant then the error should be 0, if the pixel is very oscillatory then a larger error should be expected. A more analytical justification for this is that the dominant error should come from the term in E with highest order derivatives, as is common in the analysis of PDEs.

Motivated by the examples in Section 5.2, we partition $\{\varepsilon_i\}$ into four categories of pixel sketched in Figure 5.4. These are characterised mainly by the behaviour of $\varphi^*$ on the interior of each pixel:

Type 1: $\mathrm{Lip}(\varphi^*) = \infty$, pixels where level set boundaries intersect

Type 2: $\mathrm{Lip}(\varphi^*) = O(h^{-1})$, pixels which contain multiple jumps per edge

Type 3: $\mathrm{Lip}(\varphi^*) = O(1)$, pixels on the boundaries of level sets

Type 4: $\mathrm{TV}(u) = 0$, pixels fully contained inside constant regions

Pixels of type 1 seem to be the most challenging to approximate accurately, however, in Figure 5.4 there is only one type one pixel, this is independent of $h$. This generalises; if $u^*$ has a finite number of constant regions, then there are always a finite number of pixels of type 1, therefore the total error contribution is order $h$. Similarly, type 4 pixels are trivial to discretise with no error and the works of (Chambolle and Pock, 2020; Bartels, 2020) have shown how to account for type 3. The number of pixels intersecting a level set is of order $h^{-1}$, but we achieve an error of $\varepsilon_i \lesssim h^2 \operatorname{TV}_{\omega_i}(u^*)$. Altogether, type 1, 3 and 4 pixels sum to a total error of order $h$.

It is now clear that pixels of type 2 are the only remaining obstacle for a global order $h$ convergence rate. There are order $h^{-1/2}$ pixels of type 2 but they are ignored in the analysis of Chambolle and Pock (2020); Bartels (2020), therefore we have $\varepsilon_i \lesssim h$ and can only prove $\sum_i \varepsilon_i \lesssim \sqrt{h}$. Figure 5.3 shows that this rate is sharp.

It is not clear exactly how to accurately discretise type 2 pixels. The result of Lemma 5.2.3 suggests that $u^*$ may be $\frac{1}{2}$-Hölder continuous, it might be possible to use classical methods to produce a better discretisation through the dual problem. In Section 5.3.4 we propose a more customised basis for piece-wise constant discretisation of TV. This is motivated by two graphical constructions shown in Figure 5.5. The 'thresholding' proposal is a very simple procedure which should be possible to analyse. The level set merging would preserve corner values and edge averages, but modify the level sets. Both would give the desired rate of convergence in such a simple example as in Figure 5.5 although this still needs be proven in general.

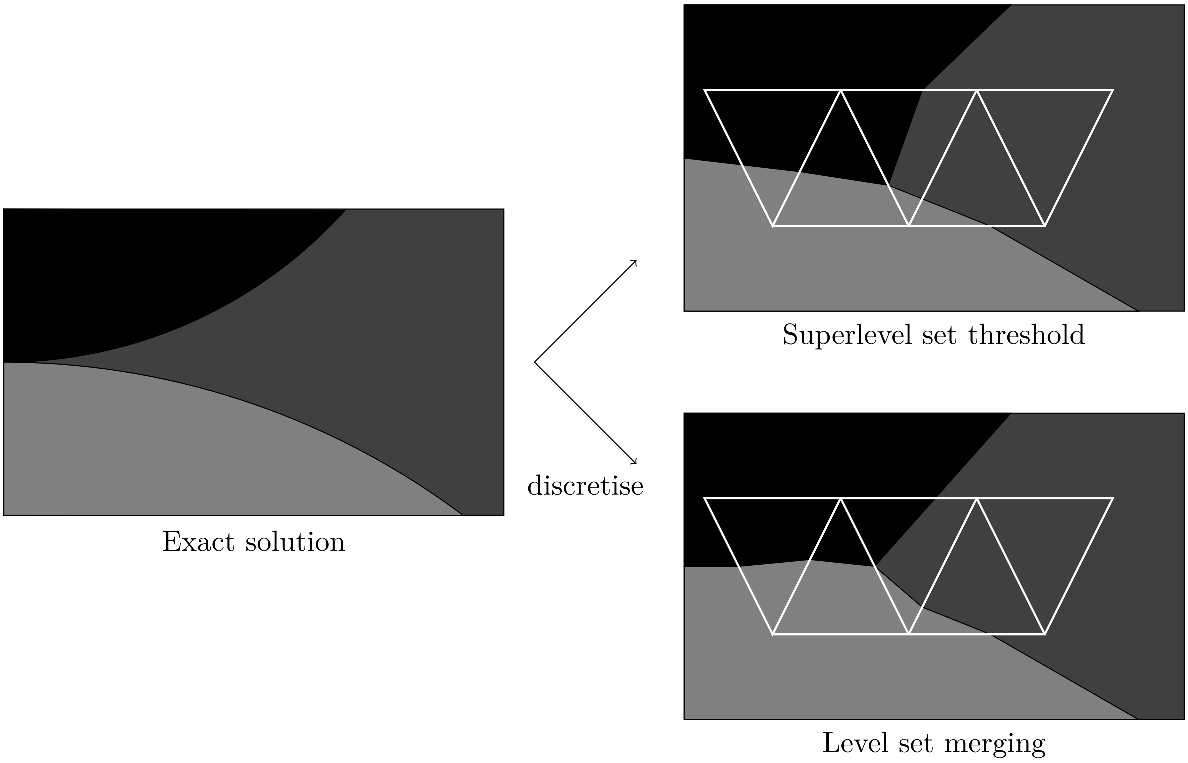


**Figure 5.5** Suggestions for high order TV discretisations. Thresholding preserves the boundary intersection points wherever possible and merging preserves the edge mean intensities.

### 5.3.3 Analytical discretisation

The aim in this subsection is to motivate that $u^*$ can be estimated using only boundary averages and corner values on each pixel. Suppose $\Omega$ is partitioned into pair-wise disjoint triangles $\mathbb{M} = \{\omega_i \text{ s.t. } i = 1, 2, \ldots\}$ with boundary $\Gamma = \{\gamma_{ij} = \partial\omega_i \cap \partial\omega_j\}$ and mesh-points $M = \{m_{ijk} = \partial\omega_i \cap \partial\omega_j \cap \partial\omega_k\}$. We will show here that if the values

$$\int_{\gamma_{ij}} u^* \qquad \text{and} \qquad u^*(m_{ijk})$$

are known exactly for each $i, j, k \leq |\mathbb{M}|$, then this is sufficient to define an approximation $u_h^*$ with the desired approximation bound.

One of the key philosophical points made by Chambolle and Pock (2020); Bartels (2020) is that E consists of an 'easy' component and a 'hard' component. If $u^*$ is in $\mathbb{BV}$, then it is smooth enough to approximate easily in $L^2$ but not in $\mathbb{BV}$. This is the result of Poincaré's inequality. If we only need an order $h$ approximation of E, then we can approximate the $L^2$ term with only pixel boundary values and the TV term performs consistent inpainting from boundary to interior. In particular, we propose the approximate energy

$$\mathrm{E}_h(u) \coloneqq \frac{1}{2}\sum_i \int_{\omega_i} \left(\eta - ⨏_{\partial\omega_i} u\right)^2 + \mu\,\mathrm{TV}(u)$$

and define the proxy total variation function

$$\mathrm{TV}(\{r_{ij}, a_{ijk}\}) = \min_u \left\{\mathrm{TV}(u) \qquad \text{s.t.} \qquad \int_{\gamma_{ij}} u = r_{ij},\ u(m_{ijk}) = a_{ijk},\ i, j, k \leq |\mathbb{M}|\right\}.$$

Hence, observe that the minimum energy can be computed:

$$\begin{aligned}\min_{u\in\mathbb{BV}(\Omega)} \mathrm{E}_h(u) &= \min_{u\in\mathbb{BV}(\Omega)} \tfrac{1}{2}\sum_i \int_{\omega_i} \left(\eta - ⨏_{\partial\omega_i} u\right)^2 + \mu\,\mathrm{TV}(u)\\ &= \min_{r_{ij}, a_{ijk}} \tfrac{1}{2}\sum_i \int_{\omega_i} \left(\eta - \frac{\sum_{ij} r_{ij}}{\sum_{ij} |\gamma_{ij}|}\right)^2 + \mu\,\mathrm{TV}(\{r_{ij}, a_{ijk}\}).\end{aligned}$$

Once the minimum energy has been computed, we can then compute

$$u_h^* = \operatorname*{argmin}_{u\in\mathbb{BV}(\Omega)} \left\{\mathrm{TV}(u) \qquad \text{s.t.} \qquad \int_{\gamma_{ij}} u = r_{ij}^*,\ u(m_{ijk}) = a_{ijk}^*,\ i, j, k \leq |\mathbb{M}|\right\}.$$

There are two key features to this approach:

- $\min_{u\in\mathbb{BV}(\Omega)} \mathrm{E}_h(u)$ can be computed on the finite dimensional space of $\{r_{ij}, a_{ijk}\}$,
- the minimiser $u_h^*$ can be visualised by solving a 'simple' TV inpainting problem,

- and we will use Poincaré's inequality to show $\mathrm{E}(u_h^*) \leq \mathrm{E}(u^*) + O(h)$.

This describes the motivation of our approach, the rest of this section is now dedicated to filling in the remaining details.

We start by showing the order $h$ approximation property. The energy can be bounded by

$$\begin{aligned}\mathrm{E}(u) - \mathrm{E}_h(u) &= \frac{1}{2}\sum_i \langle \eta - u,\ \eta - u\rangle_{\omega_i} - \left\langle \eta - \fint_{\partial\omega_i} u,\ \eta - \fint_{\partial\omega_i} u\right\rangle_{\omega_i} \\ &= \sum_i \left\langle 2\eta - u - \fint_{\partial\omega_i} u,\ \fint_{\partial\omega_i} u - u\right\rangle_{\omega_i} \\ &\leq (\|\eta\|_\infty + 2\|u\|_\infty)\sum_i \left\|\fint_{\partial\omega_i} u - u\right\|_{L^1(\omega_i)}.\end{aligned}$$

And then we introduce a TV Poincaré inequality (Ziemer, 1989, Corollary 5.12.11),

$$\sum_i \left\|\fint_{\partial\omega_i} u - u\right\|_{L^1(\omega_i)} \lesssim \sum_i \mathrm{diam}(\omega_i)\int_{\omega_i} |\nabla u|.$$

Combined, we conclude that

$$\mathrm{E}(u) - \mathrm{E}_h(u) \lesssim (\|\eta\|_\infty + 2\|u\|_\infty) h\, \mathrm{TV}(u).$$

It is known that if $\eta \in L^\infty$, then $\|u^*\|_\infty \leq \|\eta\|_\infty$ (Bartels, 2015), therefore the energy error converges with order $h$ uniformly, as required.

The final discussion of this section is whether it is computationally feasible to pursue this exact approach. Firstly, if the equation for $\mathrm{TV}(\{r_{ij}, a_{ijk}\})$ is not analytically available, then it is very hard to compute $\{r_{ij}^*, a_{ijk}^*\}$ numerically. Similarly, there is no analytical formula to compute $u_h^*$ given $\{r_{ij}^*, a_{ijk}^*\}$. For this to become a simple numerical problem, the computations would have to be pixel-wise. Analytically, the question is whether

$$\begin{aligned}&\min_{u\in\mathbb{BV}(\Omega)}\left\{\mathrm{TV}(u)\ \text{s.t.}\ \int_{\gamma_{ij}} u = r_{ij},\ u(m_{ijk}) = a_{ijk}\right\} \\ &\qquad\overset{?}{=} \sum_i \min_{u\in\mathbb{BV}(\omega_i)}\left\{\mathrm{TV}(u)\ \text{s.t.}\ \int_{\gamma_{ij}} u = r_{ij},\ u(m_{ijk}) = a_{ijk}\right\}. \qquad (5.6)\end{aligned}$$

If the answer is yes, then it means that the $\{r_{ij}, a_{ijk}\}$ discretisation behaves like a non-linear finite element basis. If the answer is no, then $u_h^*$ can still be visually rendered efficiently as a piece-wise constant function with the same order of accuracy, as guaranteed by the Poincaré inequality.

Our own conclusion is that this method is very elegant but does not lead to an efficient implementation without further modification. The important message from this subsection is that boundary values are good enough to give the desired rate, in the following section we introduce new approximations such that more of the relevant terms can be computed analytically.

### 5.3.4 Convex piecewise constant elements

We now consider forming a finite element basis on triangles which is parametrised by three (greyscale) corner values and three average edge values. This basis should ideally correspond to a convex representation for the TV value on the interior of the pixel.

#### Single positive jump case

We start with the simplest case of a single triangular pixel with a single jump, such as in Figure 5.6. First consider an example where we pin two of the corners at 0 and leave the final one free. Let $u$ be the function depicted, i.e.

$$u(\boldsymbol{x}) = \begin{cases} a & \text{on the top half} \\ 0 & \text{on the bottom half} \end{cases}, \qquad \int_{\text{left boundary}} u = r_-, \qquad \int_{\text{right boundary}} u = r_+.$$

From Section 5.3.3, there are two key properties that we would like to have analytical expressions for. The first is the map from $\{a^*, r_\pm^*\}$ to $u_h^*$, which we define to be Figure 5.6, and the second is a formula for the TV value. Computation of the TV value can be performed with the *co-area formula* (Chambolle et al., 2016).

**Lemma 5.3.1.**

$$\mathrm{TV}(u) = |r| = \left| A_\theta \begin{pmatrix} r_- \\ r_+ \end{pmatrix} \right|$$

*where*

$$A_\theta = \frac{1}{2}\begin{pmatrix} \sqrt{1-\cos\theta} + \sqrt{1+\cos\theta} & \sqrt{1-\cos\theta} - \sqrt{1+\cos\theta} \\ \sqrt{1-\cos\theta} - \sqrt{1+\cos\theta} & \sqrt{1-\cos\theta} + \sqrt{1+\cos\theta} \end{pmatrix}.$$

*Proof.* The total variation of $u$ is the length of the interior jump times scaled by jump height (Chambolle et al., 2016), in Figure 5.6 we assume $r$ already denotes length scaled by $a$. The

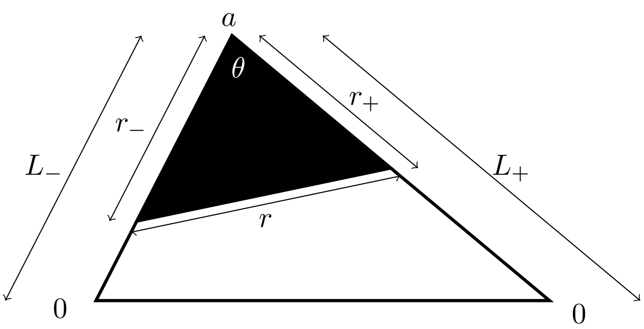


**Figure 5.6** Simple triangle discretisation with single jump.

remaining formula is just the cosine rule:

$$\frac{r^2}{a^2} = \frac{r_-^2}{a^2} + \frac{r_+^2}{a^2} - 2\cos(\theta)\frac{r_-r_+}{a^2}.$$

If we assert that $\mathrm{TV}(u)$ is a norm in the form proposed in the lemma, then we get

$$\begin{aligned}\left|A_\theta\begin{pmatrix}r_-\\ r_+\end{pmatrix}\right|^2 &= (A_{11}r_- + A_{12}r_+)^2 + (A_{21}r_- + A_{22}r_+)^2\\ &= (A_{11}^2 + A_{21}^2)r_-^2 + (A_{12}^2 + A_{22}^2)r_+^2 + 2(A_{11}A_{12} + A_{21}A_{22})r_-r_+\\ &= \tfrac{1}{2}(1-\cos\theta + 1 + \cos\theta)(r_-^2 + r_+^2) + (1-\cos\theta - 1 - \cos\theta)r_-r_+\end{aligned}$$

as required. □

The key result of this lemma is that the function is still convex and easily computable. In the notation of Section 5.3.3, we can write the TV functional as

$$\mathrm{TV}(\{(a,0,0),(r_-,r_+,0)\}) = \left|A_\theta\begin{pmatrix}r_-\\ r_+\end{pmatrix}\right|.$$

The formula is very simple but the dependence on $a$ has become an implicit feasibility constraint. The total variation is given by this formula so long as

$$r_\pm = s_\pm L_\pm a \text{ for some } s_\pm \in [0,1].$$

This unfortunately conflicts with our final aim from Section 5.3.3; the inequality is a non-convex constraint. This can be confirmed with a simple example, let $L_\pm = 1$ and denote pairs $(r,a)$. Both $(1,1)$ and $(-\frac{1}{2},-1)$ are feasible but their midpoint $(\frac{1}{4},0)$ is not feasible. This is a surprising limitation, but can be overcome by a simple non-negativity constraint:

$$0 \leq r_\pm \leq L_\pm a.$$

This feasibility condition is now convex and can be extended to $a \in \mathbb{R}$ by doubling the number of variables and writing, for example, $a = \max(0,a) - \max(0,-a)$ as the difference of two new non-negative variables.

Recall that, in 1D, we can write any function as the difference of two non-decreasing functions. This appears to be the natural mindset for convexifying our adaptive piecewise constant discretisation.

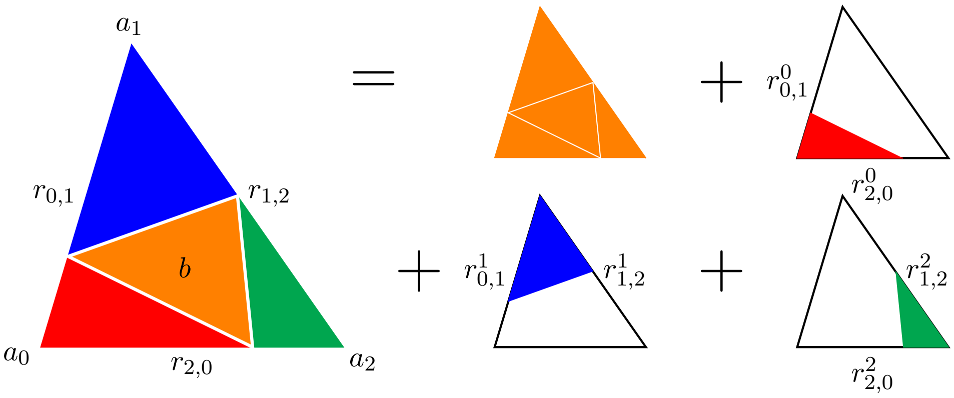


**Figure 5.7** Full triangle discretisation with multiple jumps. The given parameters are $a_i$, and $r_{i,i+1}$. $b$ is free to be chosen. One pixel is divided into four component parts.

### General elements

To generalise the element shown in Figure 5.6, we decompose each triangle into four components; one constant and three corner jumps (i.e. one per corner, see Figure 5.7). The corner and edge values are considered the main parameters for optimisation, as motivated in Section 5.3.3. The final constant term, labelled $b$, can either be another optimisation parameter or a scalar with fixed dependence on the original parameters. The role of $b$ is not important and will not be discussed further in this section.

As before, we define the map from $\{a_i^*, r_{i,j}^*\}$ to $u_h^*$ by Figure 5.7, and proceed to compute an analytical form for the total variation value and corresponding feasibility constraints.

To compute the TV value, we need to apply Lemma 5.3.1 once per corner. To do this, we need to split each edge term into two components, as in Figure 5.7. For example, $r_{0,1}$ is split into an $a_0$ component and a $a_1$ component such that

$$r_{0,1} = r^0_{0,1} + r^1_{0,1} + L_{0,1}b$$

where $L_{0,1}$ is the side length between the two corners. The TV value can then be expressed in the form

$$\mathrm{TV}(\{a_i\}, \{r_{i,i+1}\}) = \sum_{i\in\{0,1,2\}} \left| A_i \begin{pmatrix} r^i_{i,i+1} \\ r^i_{i-1,i} \end{pmatrix} \right|$$

where $A_i$ are appropriately chosen matrices depending on the interior angle at corner $i$ and we work with $i \in \mathbb{N} \bmod 3$. This function is simple and convex, but requires a large dimensionality lifting:

- We start (left-hand of Figure 5.7) with six variables in $\mathbb{R}$, and three non-convex inequalities.
- Each pixel is decomposed into 12 variables in $\mathbb{R}_{\geq 0}$, three linear equalities, and six non-convex inequalities.

- The constraints are convexified into 24 variables in $\mathbb{R}_{\geq 0}$, six linear equalities, and 12 linear inequalities.
- Substituting the linear constraints leaves 18 variables in $\mathbb{R}_{\geq 0}$ and 12 linear inequalities.

While it is technically possible to implement this discretisation, this represents a very large increase in computational complexity which can hopefully be reduced before attempting an implementation.

## 5.4 Discussion

The result of Section 5.2 demonstrates that there is still scope for improvement in the discretisation for TV optimisation. Introducing this new more challenging example provides a new analytical problem to benchmark against.

It appears clear now that the most challenging areas to discretise are the neighbourhoods of intersections of level sets. Some well-known examples are circles which result in level sets with constant curvature after denoising, and squares which lead to piecewise constant curvature. Any $u^*$ with level sets of piecewise constant curvature would satisfy the same error rates. If a 'worse example' exists, then the level sets must have vanishing curvature at the point where they intersect. It is unclear whether this is possible.

On the numerical side, the ideas of Section 5.3.3 look very elegant, but the concrete suggestion at the end of Section 5.3.4 leaves a lot to be desired. If either formula in (5.6) could be expanded analytically, then this could allow direct minimisation of (5.1). Instead, we introduce a very bloated lifting of the original discretisation which results in a simple formula. If the number of parameters could be at least halved then it would begin to look like a practical proposal.

One parameter which hasn't been discussed yet is the role of $b$ in Figure 5.7. It could be left as a variable or there are two natural choices for simplifying the problem. It can be shown that

$$b = \operatorname{median}(a_i)$$

is the minimal TV choice. This exactly agrees with the sparse-gradient interpretation of TV. Another intuitive choice is

$$b = \frac{\sum r_{i,i+1} L_{i,i+1}}{\sum L_{i,i+1}} = \fint_{\partial\omega_i} u$$

which is the value used in the data fidelity. The first choice has potential to achieve one of the 'continuous optimum' forms given in (5.6) although there is no proof. Other than this, there are no particular analytical guarantees corresponding to the choice of $b$.

# Chapter 6

# Reconstruction with a Gaussian Dictionary

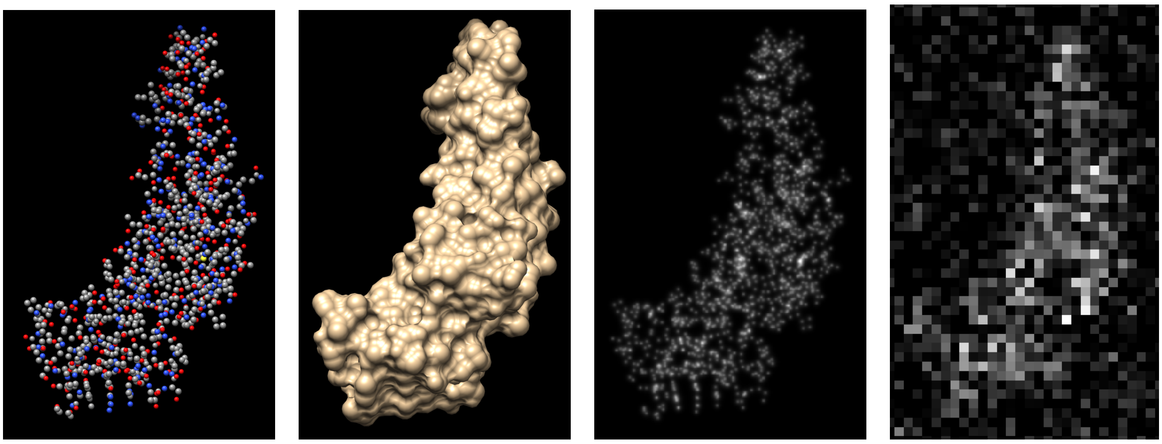
Single particle cryo-electron microscopy is an exciting modality with the ability to perform atomic resolution reconstructions of biological samples. As an inverse problem, we can describe the task as reconstruction of atomic resolution densities from a very large amount of very poor X-ray data. The full pipeline of single particle analysis is very long and much more complex than the component considered in this chapter. In particular, data pre-processing requires specialised algorithms for: deconvolution, segmentation, clustering, registration, and super-resolution. On top of this, the tomographic component of the inverse problem is also non-standard as the imaging orientations are unknown. For further details see Moriya et al. (2017); Righetto et al. (2019).

In this chapter we focus on the final stage of the reconstruction, where an atomic model is computed from noisy X-ray data (with known orientations). An example is pictured in Figure 6.1a. The data shall consist of around $10^4$ X-ray projections of the quality seen in Figure 6.1d and the desired output is seen in Figure 6.1a. We will use the term *atomic reconstruction* to refer to any discrete reconstruction which identifies a 3D centre of each atom in the volume.

For brevity of notation, we will say that $u/v$ denote atomic/volume reconstructions (Figures 6.1a and 6.1b) respectively while $\boldsymbol{\eta}$ represents the raw data (Figure 6.1d).

The *Protein Data Bank*[1] is a large database with over 5000 atomic reconstructions attributed to single particle analysis. The standard reconstruction method is to first reconstruct the 3D scattering potential ($\boldsymbol{\eta} \mapsto v$) then fit atomic potentials to that 3D reconstruction ($v \mapsto u$). We will refer to this as the *sequential method.* The initial reconstruction is typically very simple, for example classic Tikhonov regularisation (Donati et al., 2018; Zivanov et al., 2019). The atomic registration is then performed as a highly customised Gaussian mixture decomposition (Murshudov et al., 2011; Liebschner et al., 2019) (see Section 6.1 for a concrete formulation).

[1] https://www.rcsb.org/



(a) Atomic sample (b) Physical sample (c) Analytical data (d) Simulated data

**Figure 6.1** Different representations of a protein. (a) shows the 3D atomic representation, colour encodes the type of atom. (b) shows an isosurface of the 3D scattering potential. (c) shows the analytical X-ray transform of the sample in (b). In practical examples data is coarse and noisy as demonstrated in (d). 3D renders from UCSF Chimera (Pettersen et al., 2004).

The sequential method has proved highly successful, but we would also like to perform atomic fitting directly to the data ($\boldsymbol{\eta} \mapsto u$). This approach received a lot of interest in 2015 (Joubert and Habeck, 2015; Goris et al., 2015; Xu et al., 2015), although has so far failed to gain mainstream popularity.

In this chapter we will make some numerical observations on how the quality of atomic reconstruction vary with respect to various parameter choices. The main questions we ask are:

- The impact of problem formulation. Is it more accurate to fit directly ($\boldsymbol{\eta} \mapsto u$) or sequentially ($\boldsymbol{\eta} \mapsto v \mapsto u$)?
- The impact of optimisation model. We will define our ideal reconstruction as the minimiser of a function. Can we add constraints to the optimisation to find more accurate reconstructions?
- The impact of numerical methods. In Section 6.1 we will see that atomic fitting is a non-convex problem. Is there a numerical scheme which consistently finds better minimisers?

We emphasise that the results shown here are not a benchmark of several state-of-the-art approaches. In each approach we use the simplest implementation to indicate the characteristic behaviour of that approach.

## 6.1 Algebraic formulation

Raw data will be denoted $\boldsymbol{\eta}$, in the form seen in Figure 6.1d. The associated forward map is a subsampled X-ray transform $\mathcal{R}\colon L^1([0,1]^3) \to \mathbb{R}^{m_1\times m_2^2}$

$$\mathcal{R}[v](\boldsymbol{\theta}, \boldsymbol{x}) = \int_{\mathbb{R}} v(\boldsymbol{x}+t\boldsymbol{\theta})dt$$

for $\boldsymbol{\theta} \in \Theta \subset \mathbb{S}^2$ and $\boldsymbol{x}$ on an appropriate square grid. We will consider atomic reconstructions $u$ to be indexed tuples of the form

$$u_i = (\alpha_i,\ \boldsymbol{x}_i,\ R_i) \in \mathbb{R}\times\mathbb{R}^3\times\mathbb{R}^{3\times 3}, \qquad i = 1,\dots,N.$$

This represents the Gaussian density

$$v(\boldsymbol{x}) = \sum_{i=1}^N \alpha_i \exp\left(-\tfrac{1}{2}|R_i(\boldsymbol{x}-\boldsymbol{x}_i)|^2\right).$$

The average number of atoms for samples in the protein data bank is $N \approx 3\cdot 10^3$ although more interesting complexes are at least a factor of 10 larger. For example, a single spike on the surface of SARS-Cov-2 has around $2\cdot 10^4$ atoms[2].

We will consider three reconstructions:

- the volume reconstruction

$$v^* = \operatorname{argmin}\left\{\|v\|_2^2 \text{ s. t. } \mathcal{R}v = \boldsymbol{\eta}\right\}, \tag{6.1}$$

- the sequential reconstruction

$$u_S^* = \operatorname{argmin}\left\|\sum_{i=1}^N \alpha_i \exp\left(-\tfrac{1}{2}|R_i(\cdot-\boldsymbol{x}_i)|^2\right) - v^*\right\|_2^2, \tag{6.2}$$

- and the direct reconstruction

$$u_D^* = \operatorname{argmin}\left\|\sum_{i=1}^N \alpha_i\mathcal{R}\left[\exp\left(-\tfrac{1}{2}|R_i(\cdot-\boldsymbol{x}_i)|^2\right)\right] - \boldsymbol{\eta}\right\|_2^2. \tag{6.3}$$

The optimisation problem (6.2) is generally referred to as *Gaussian mixture decomposition.* One reason Gaussians are useful as a basis for X-ray inverse problems is because the forward

[2] https://www.rcsb.org/structure/6X6P

model is easily computable:

$$\begin{aligned}\mathcal{R}\left[\exp\left(-\tfrac{1}{2}|R_i(\cdot-\boldsymbol{x}_i)|^2\right)\right](\boldsymbol{\theta},\boldsymbol{x}) &= \int_{\mathbb{R}}\exp\left(-\tfrac{1}{2}|R_i(\boldsymbol{x}+t\boldsymbol{\theta}-\boldsymbol{x}_i)|^2\right)dt\\ &= \int_{\mathbb{R}}\exp\left(-\tfrac{1}{2}|R_i(\boldsymbol{x}-\boldsymbol{x}_i)+tR_i\boldsymbol{\theta}|^2\right)dt\\ &= \exp\left(-\tfrac{1}{2}|\Pi_{R_i\boldsymbol{\theta}}R_i(\boldsymbol{x}-\boldsymbol{x}_i)|^2\right)\int_{\mathbb{R}}\exp\left(-\tfrac{t^2}{2}|R_i\boldsymbol{\theta}|^2\right)dt\\ &= \frac{\sqrt{2\pi}}{|R_i\boldsymbol{\theta}|}\exp\left(\tfrac{1}{2}\left(\frac{R_i\boldsymbol{\theta}}{|R_i\boldsymbol{\theta}|}\boldsymbol{\cdot}R_i(\boldsymbol{x}-\boldsymbol{x}_i)\right)^2-\tfrac{1}{2}|R_i(\boldsymbol{x}-\boldsymbol{x}_i)|^2\right)\end{aligned}$$

where $\Pi_{R_i\boldsymbol{\theta}}$ is the projection onto the orthogonal complement of $R_i\boldsymbol{\theta}$. The take-home of this formula is that it is very simple to analytically evaluate the X-ray transform of a Gaussian kernel, even when it is anisotropic. If the kernel is isotropic ($R_i\in\mathbb{R}$), then it simplifies even further to

$$\mathcal{R}\left[\exp\left(-\tfrac{R_i^2}{2}|\cdot-\boldsymbol{x}_i|^2\right)\right](\boldsymbol{\theta},\boldsymbol{x}) = \frac{\sqrt{2\pi}}{R_i}\exp\left(\tfrac{R_i^2}{2}\left(\boldsymbol{\theta}\boldsymbol{\cdot}(\boldsymbol{x}-\boldsymbol{x}_i)\right)^2-\tfrac{R_i^2}{2}|\boldsymbol{x}-\boldsymbol{x}_i|^2\right).$$

If the $\frac{1}{R_i}$ scaling is absorbed into $\alpha_i$, then the whole function becomes smooth and the numerical properties of Equations (6.2) and (6.3) look identical.

## 6.2 Motivation for direct approach

There are several reasons to anticipate that the direct approach of (6.3), fitting the atomic reconstruction to the raw data, might be better than sequentially solving the least squares and then atomic decomposition (i.e. (6.1) then (6.2)).

The key philosophical problem with the sequential approach is that the final reconstruction is not verified against the raw data. Information can only be lost during the reconstruction, therefore it should not be possible to achieve a better result using only $v^*$ and not referring back to $\boldsymbol{\eta}$.

Computation of $v^*$ in (6.1) relies on computing a least squares minimiser on a particular discretised grid. The least squares solution of an ill-posed inverse problem is known to amplify noise in the data, which will add to any new discretisation artifacts. The computation of $u_S^*$ must overcome each of these sources of noise in $v^*$.

The direct reconstruction has fewer parameters which can affect the quality of reconstruction. The only pixel discretisation is the data, determined by the microscope, and the choice of atomic discretisation is equivalent in both approaches. The prior that each atom should look like a Gaussian is very strong and should be more than sufficient to perform the required denoising and super-resolution tasks without choosing an intermediate prior for $v^*$.

## 6.3 Numerical methods

Each of the optimisation formulae (6.2) and (6.3) are smooth non-convex problems, let E denote the chosen energy function. In non-convex optimisation different numerical schemes give different results and so we would like to compare three different methods. In Section 6.4 we will also need to include a convex constraint such that $u_i \in C$ for each $i$.

Our baseline numerical method is block gradient descent with backtracking. In particular, for one descent step (fixed $n$) we sequentially compute for each $i$

$$u_i^{n+1} = \operatorname*{argmin}_{u\in C} \nabla_i \operatorname{E}(u_1^{n+1}, \ldots, u_{i-1}^{n+1}, u_i^n, u_{i+1}^n, \ldots, u_N^n) \boldsymbol{\cdot} (u - u_i^n) + \frac{1}{2\tau_i^n} \|u - u_i^n\|_2^2$$

where $\tau_i^n > 0$ is the stepsize. If $\operatorname{E}(u_i^{n+1}) > \operatorname{E}(u_i^n)$ then we revert to $u_i^{n+1} = u_i^n$ and set $\tau_i^{n+1} = 0.9\tau_i^n$.

Our second method is a block Newton descent which is better able to capture the local curvature of E. The formula is given by

$$u_i^{n+1} = \operatorname{argmin}_{u\in C} \nabla_i \operatorname{E}(\ldots, u_{i-1}^{n+1}, u_i^n, \ldots) \boldsymbol{\cdot} u + \left(\left[\tfrac{1}{2\tau_i^n} + \nabla_{i,i}^2 \operatorname{E}(\ldots, u_{i-1}^{n+1}, u_i^n, \ldots)\right] (u - u_i^n)\right) \boldsymbol{\cdot} (u - u_i^n).$$

This is a little more complex than gradient descent but still efficiently computable. Again, if $\operatorname{E}(u_i^{n+1}) > \operatorname{E}(u_i^n)$ then we revert to $u_i^{n+1} = u_i^n$ and set $\tau_i^{n+1} = 0.9\tau_i^n$.

Our final method is a stochastic variant of the Newton descent algorithm. In particular, we use a *Metropolis-Hastings* (MH) algorithm referred to as *PMH2* by Dahlin et al. (2015). First we compute the exact Newton step,

$$\widehat{u} = \operatorname*{argmin}_{u\in C} \nabla_i \operatorname{E}(\ldots, u_{i-1}^{n+1}, u_i^n, \ldots) \boldsymbol{\cdot} u + \left(\left[\tfrac{1}{2\tau} + \nabla_{i,i}^2 \operatorname{E}(\ldots, u_{i-1}^{n+1}, u_i^n, \ldots)\right] (u - u_i^n)\right) \boldsymbol{\cdot} (u - u_i^n),$$

then sample a random *candidate*

$$u_i^{n+1} \sim \mathcal{N}(\widehat{u}, \nabla_{i,i}^2 \operatorname{E}(\ldots, u_{i-1}^{n+1}, u_i^n, \ldots)^{-1}).$$

If $\operatorname{E}(u_i^{n+1}) \ll \operatorname{E}(u_i^n)$ then the candidate is accepted with high probability, otherwise the proposal can be rejected leading to $u_i^{n+1} = u_i^n$. We will not go further into the acceptance procedure here, we use the standard Metropolis-Hastings technique which is written explicitly in Chib and Greenberg (1995). The idea is that the acceptance strategy should optimally balance 'exploitation' (locally minimising E) and 'exploration' (finding global minimisers). The idea of an exploration/exploitation trade-off is very standard in non-convex optimisation, see Gittins et al. (2011) for a more detailed explanation.

## 6.4 Numerical results

For numerical comparisons we use a tobacco-mosaic virus[3] which is a small structure with 1206 atoms. The data was simulated using the ASTRA toolbox (Van Aarle et al., 2016) from a 0.1 Å grid to $10^4$ projections at resolution 1.5 Å on a grid of size $22 \times 22$. $v^*$ was computed at resolution 0.2 Å on a grid of size $104 \times 136 \times 148$. The resolutions were chosen to be physically realistic. As a rule of thumb, an atom is approximately 1 Å in diameter.

Our default constraint set is

$$C = \left\{(\alpha, \boldsymbol{x}, R) \in \mathbb{R} \times \mathbb{R}^3 \times \mathbb{R}^{3\times 3} \text{ s. t. } \alpha \geq 0, \text{ eigenvalues of } R \text{ are between } 0.3\,\text{Å}^{-1} \text{ and } 1\,\text{Å}^{-1}\right\}.$$

Due to the number of reconstructions that we compare, each component is assigned a key. The optimisation schemes we use are:

GD: Gradient descent constrained to $C$

N: Newton descent constrained to $C$

MH: Metropolis-Hastings constrained to $C$

I: Isotropic atoms constrained to $C$ such that $R_i \in \mathbb{R}$, optimised with the Newton algorithm

R: Fixed radius atoms with $\alpha_i \geq 0$ and $R_i = 0.3\,\text{Å}^{-1}$, optimised with the Newton algorithm

We run each of these optimisation algorithms on the direct/sequential formulations with two different levels of noise and two initialisations:

D/S: Direct/Sequential optimisation.

$a/a^+$: The dataset is either exact or corrupted with Gaussian white noise of variance 12 % of the maximum data intensity (signal-to-noise ratio of 0.39).

$b/b^+$: The first initialisation is close to the true solution, each atom is displaced by a uniform random perturbation in the range $[-1, 1]^2$. In particular, each atom is still 'touching' one of the Gaussians in the initialisation. The second initialisation is uniformly random over the sample volume.

Figures 6.2 to 6.6 report the median resolution of each reconstruction. This is defined as the median over all atoms of the minimal distance to a Gaussian in the reconstruction. In particular, if the resolution is quoted at >1.5 Å, then at least 50 % of atoms are at least 1.5 Å from the nearest Gaussian. This is a significant threshold because it shows that the atomic fitting has failed to reconstruct to the resolution of the data, which is also at 1.5 Å.

Each figure shows the full distribution of errors over each reconstruction type although they should be used mainly to identify trends. It is not necessary to decode the exact parameters of each reconstruction although a key is provided in each caption.

[3] https://www.rcsb.org/structure/6I5A

**Comparison of algorithms** In Figure 6.2 we compare the performance of each optimisation algorithm. The models in (6.2) and (6.3) define functions such that the minimisers should be close to the true atomic structure. We are therefore interested in two quantities; the final function value and the median resolution. The energy is directly computed from the data, whereas the resolution is more physically meaningful and requires knowledge of the ground truth solution. In terms of energy minimisation, we see that Newton is almost always the most powerful method. The link between function value has and resolution is less clear. For both gradient and Newton descent there is even a negative correlation; lower energies correspond to worse reconstructions. This is particularly noticeable in the Newton method with poor initialisation; the energy is good but the resolution is very poor. On the other hand, the MH method achieves both the best resolution and lowest energy.

At this point we note that the MH method is not designed to converge to minimisers, so it is not completely fair to compare each method equally in this regard. The convergence behaviour of MH can be summarised to the statement that $\mathrm{E}(u^n)$ will be small with high probability for large $n$. A stronger and more precise statement is made by Chib and Greenberg (1995). In the context of atomic fitting, we make the assumption that the hardest component of the optimisation problem is finding global as opposed to local minimisers. This justifies the inclusion of MH as an optimisation scheme because it has good global guarantees, whereas the other monotone descent schemes only offer local guarantees. The results in Figure 6.2 confirm that the local/global trade-off of MH is sufficient for competitive results in this application without further modification.

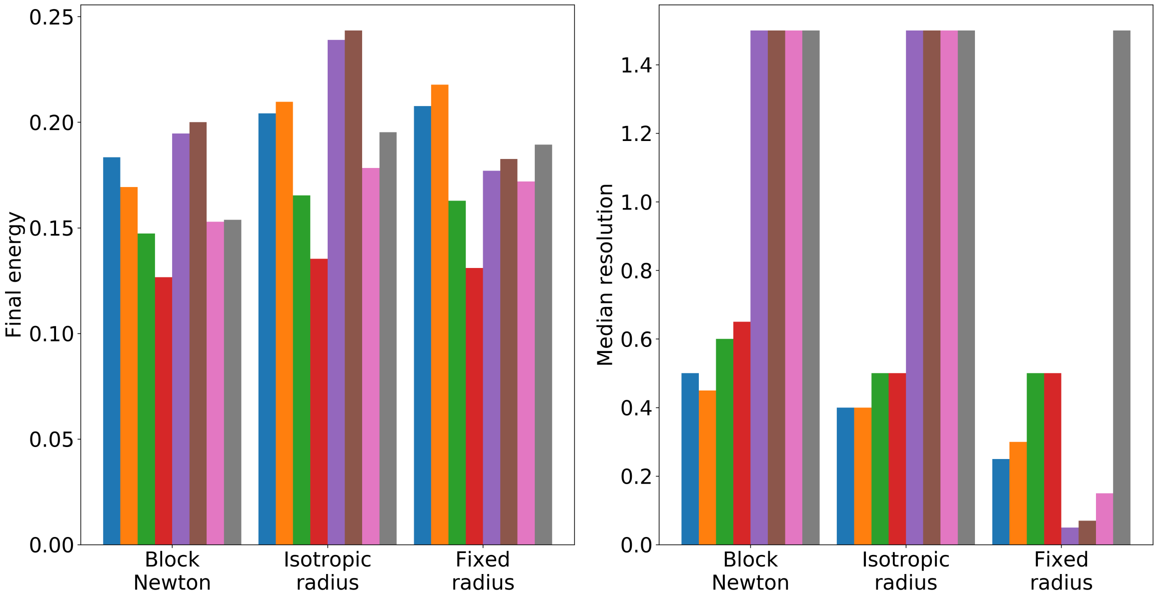


**Figure 6.3** Comparison of different constraints with the same optimisation scheme. In order, bars represent D$ab$, D$a^+b$, D$ab^+$, D$a^+b^+$, S$ab$, S$a^+b$, S$ab^+$, S$a^+b^+$. Fixing the radius always improves resolution of reconstruction, even when the energy becomes worse.

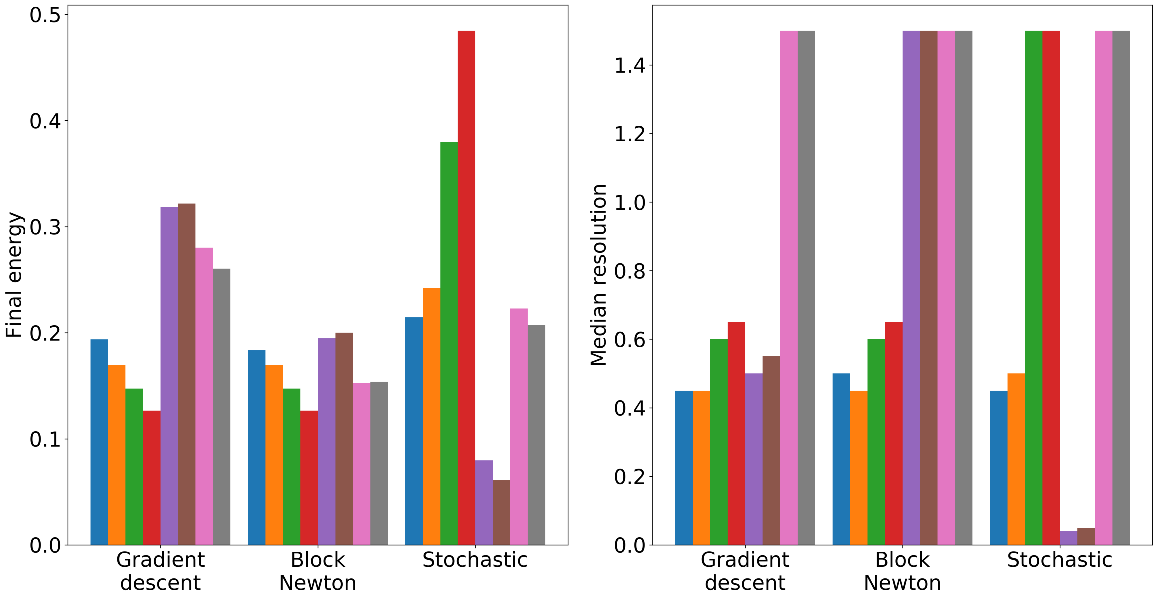


**Figure 6.2** Comparison of different algorithms for the same energy. In order, bars represent D$ab$, D$a^+b$, D$ab^+$, D$a^+b^+$, S$ab$, S$a^+b$, S$ab^+$, S$a^+b^+$. There is a poor correlation between energy and resolution. Newton descent is never worse than gradient descent.

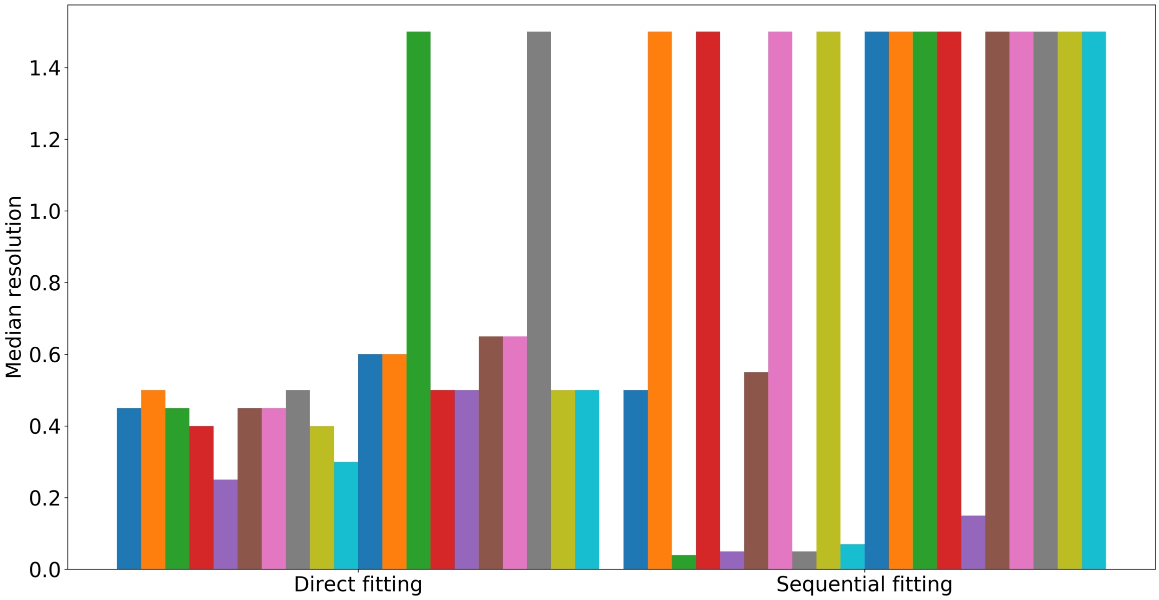


**Figure 6.4** Comparison of direct and sequential optimisation. The bars are ordered in blocks of five indicating optimisation scheme (GD, N, MH, I, R). The blocks are ordered $ab$, $a^+b$, $ab^+$, $a^+b^+$. Direct is more robust but sequential has better peak performance.

**Comparison of constraints** In Figure 6.3 we explore the impact of adding extra constraints to the Gaussian parameters. We use the Newton scheme in each case as it was the most powerful in Figure 6.2. We observe the lack of correlation between function value and resolution as seen in Figure 6.2. In these experiments the fixed radius was held at $R_i = 0.3\,\text{Å}^{-1}$ which corresponds to a very large atom. Despite this poor physical motivation, fixing a large radius appeared to make the reconstructions more physically accurate. This is particularly apparent in the sequential optimisation.

**Comparison of direct/sequential** Figure 6.4 compares direct and sequential optimisation. We observe that sequential fitting achieves the peak optimal resolution but is also much more likely to fail to find half of the atoms. In 'easy' scenarios initialised close to the exact sample, both the Metropolis-Hastings and fixed radius optimisation schemes are capable of very high accuracy; approximately five times better than the best direct reconstruction. On the other hand, direct reconstruction is much more robust at finding the majority of atoms; 10% failure vs 65%.

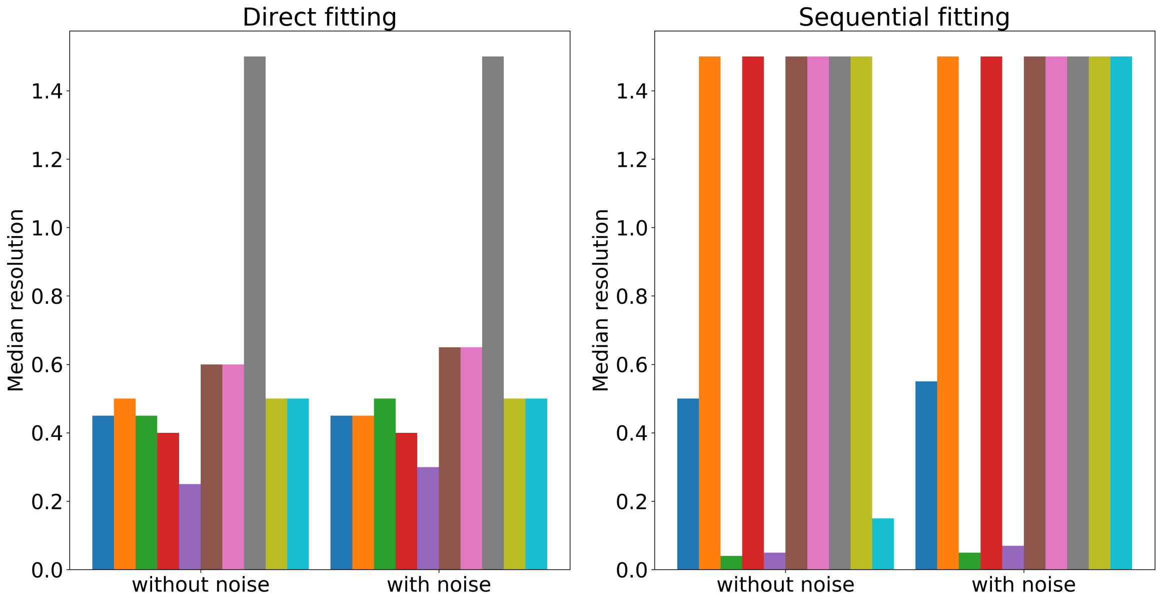


**Figure 6.5** Comparison of different noise levels. In order, bars represent GD$b$, N$b$, MH$b$, I$b$, R$b$, GD$b^+$, N$b^+$, MH$b^+$, I$b^+$, R$b^+$. Every method is very robust to noise.

**Comparison of noise levels** In Figure 6.5 we see that adding a large amount of noise has no practically observable impact on accuracy in any example. This confirms that the atomic fitting procedure is very robust to noise.

**Comparison of initialisation quality** The final comparison in Figure 6.6 tests how the quality of the initialisation affects the quality of the reconstruction. The impact is much clearer than with noise, almost all direct methods retain reasonable accuracy while almost all sequential methods fail with the uniform random initialisation. The only exceptions to this for direct optimisation are the Metropolis-Hastings methods, although this is likely to be an indicator of slow convergence rather than a more fundamental limitation of the method. The only sequential optimisation capable of finding 50% of the atoms with a bad initialisation is when the Gaussian radius is fixed.

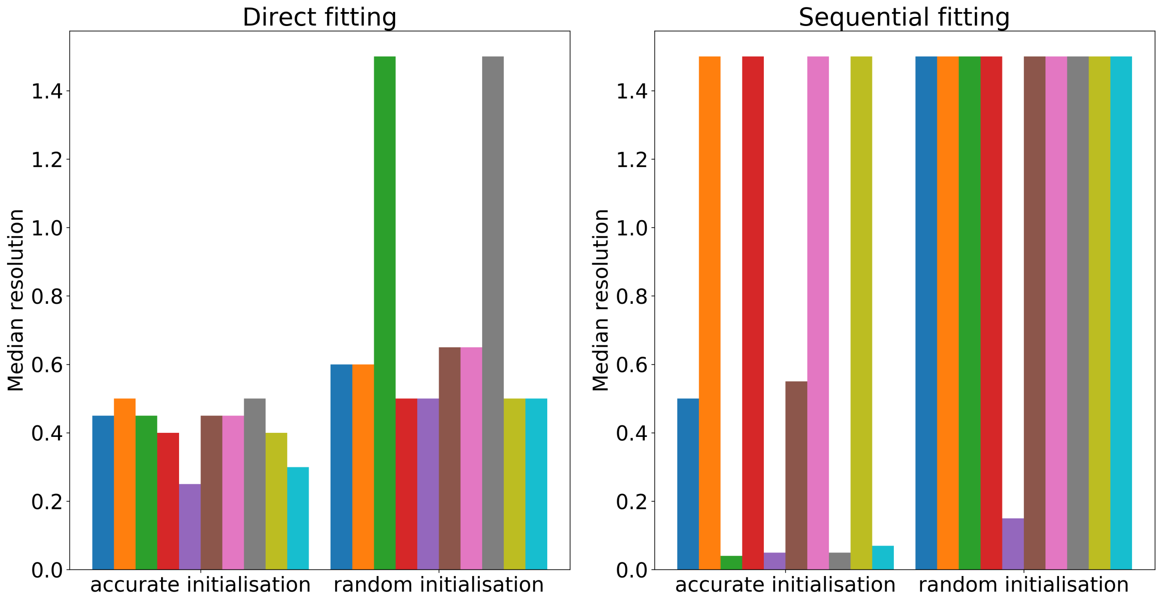


**Figure 6.6** Comparison of different initialisations. In order, bars represent GD$a$, N$a$, MH$a$, I$a$, R$a$, GD$a^+$, N$a^+$, MH$a^+$, I$a^+$, R$a^+$. Poor initialisation always reduces final resolution but the sequential optimisation is much more sensitive.

## 6.5 Discussion

Section 6.4 presents some interesting initial observations although much more thorough studies would be needed to make confident conclusions. In particular, the current experiments should be repeated with several random instances to determine average behaviour. We hope this effect is somewhat replicated by averaging over the 1000 atoms although these errors are not independent. We must also show that our observations are consistent with large proteins, for example using a more standard benchmark problem in the range 12,000 to 50,000 atoms such as recommended by Kim et al. (2018).

A final limitation of the current study is that all tested implementations are the simplest within their class. This gives us a baseline understanding of the characteristics of each approach but practical performance also depends on the quality of current state-of-the-art algorithms. A full review should also take this factor into account.

Aside from these caveats, what we have seen suggests that there are some common trends for the reconstruction of atomic models from X-ray tomography data. One clear conclusion from Figures 6.2 and 6.3 is that the data error cannot be relied upon as a fine-scale proxy for physical accuracy. We have seen several reconstructions which fit the data well but have poor accuracy, and the same situation in reverse. Figure 6.5 confirms that all atomic reconstructions are very robust to noise. Further tests should be carried out with non-Gaussian noise but the current results are very encouraging. Figure 6.6 suggests that direct reconstructions are also robust to initialisation but sequential are not. In particular, the direct reconstruction resolution increases only from 0.3 Å to 0.6 Å. From an applied viewpoint, there is little gained by this scale of improvement therefore a direct reconstruction could be used without developing any more advanced initialisation strategy, which is necessary for the indirect optimisation.

One of the main comparisons we wished to make was between direct and indirect optimisation. Figure 6.4 clearly reinforces the conclusion of Figure 6.6 that direct fitting is more stable but sequential fitting has better peak performance. This suggests that the optimal strategy is to use the direct method as an initialisation and the sequential for final refinement; the opposite to our initial prediction in Section 6.2. The key difference may be the locality properties of the forward operator in each method. Atomic charge decays exponentially therefore the sequential problem (6.2) of fitting Gaussians to $v^*$ depends only on local values of $v^*$ within a distance of approximately 2 Å from the centre of the Gaussian. If two atoms/Gaussians are far apart then they do not interact in the optimisation. On the other hand, in the direct approach two atoms overlap in data space whenever there exists a line of direction $\boldsymbol{\theta} \in \Theta$ which passes through each atom. As $\Theta$ is a very large set, it is very likely that every atom and every Gaussian in the reconstruction interact at some pixel in $\boldsymbol{\eta}$.

The local nature of the sequential forward map means that it is very hard to find atoms with gradient based methods, if the Gaussian does not overlap the atom then there is no attraction in the gradient (or at least it is exponentially small). On the other hand, the global properties of the direct forward map means that every Gaussian feels some attraction to every atom. This is good for finding new atoms but possibly also explains the poor local resolution. Each Gaussian is trying to resolve every atom at once rather than focussing on the nearest. This heuristic is so far consistent and sufficient to explain the observed behaviour.

Figure 6.2 shows that Newton descent can find lower energy minimisers than gradient descent. This is interesting because it indicates that Newton descent is 'better' rather than simply faster in this setting. When the non-convexity becomes more apparent, i.e. in the sequential optimisation, then only the stochastic scheme is capable of performing well. This aligns with the initial motivation for including a MH algorithm for comparison; reconstructions change quickly while the energy is high then slow down in low energy regions.

Figure 6.3 showed that adding further constraints to the optimisation can greatly improve results. The most surprising observation was that fixing an un-physically large radius is most effective at improving resolution. From a numerical point of view, this is very convenient because it removes many parameters from optimisation.

# Chapter 7

# Discussion and Outlook

This thesis has contributed to several areas of the mathematics of electron tomography including physical modelling, mathematical modelling, and mathematical optimisation. Each chapter has been self-contained thus far and so we will continue the discussion in the same way.

## 7.1 Limited angle tomography

In Chapter 2 we proposed a new reconstruction model for limited angle tomography and a new numerical algorithm for computing minimisers of this model. This model showed consistent qualitative advantages over the previous state of the art method.

### 7.1.1 Reconstruction model

One of the realisations during this project for myself was how hard it is to categorically ‘beat’ the standard TV reconstruction, despite how easy it was to ‘break’ it initially. With up to about 25 % missing data, the reconstruction is unaffected by the structure of the missing data. Going past this percentage, the Fourier slice theorem tells us that there are ‘visible’ and ‘invisible’ structures, standard TV achieves near perfect reconstruction in the visible component while noticeably failing to reproduce the invisible component reliably.

The proposed model is much better at recovering large scale features in the invisible structures but fails to preserve the high quality of the visible. In my eyes, this is the major limitation of our proposed directional TV. The limited angle application is a clear example where the TV functional does not directly encourage sparse jumps, which is why we could consistently ‘break’ the TV reconstruction, e.g. Figure 2.2. In the regions with missing data there is a clear smoothing effect which does not correspond to sparsity. Once we added in the optimisation of the directional component, this blurring then spread into the visible domain. For an inpainting problem, this is precisely the behaviour one does not want outside of the inpainting domain. It is unclear how to overcome this problem at the current time. Clearly the TV prior is not robust enough, a non-convex sparse gradient penalty may improve reconstructions although this would still have to overcome the same local minimiser difficulties of our own work.

In the years surrounding this publication there were several new proposals for spatially adaptive TV regularisation, mainly focussing on the direct measurement case. Work by Hintermüller et al. (2017, 2018) framed denoising as a bi-level optimisation problem where one uses knowledge of the noise to adapt the (isotropic) weighting of TV. In other work, such as by Calatroni et al. (2019), anisotropy is detected in noisy images (but does not adapt during the optimisation) and a $p$-norm with $p < 1$ is used to encourage sparsity more strongly in reconstructions. Both of these approaches have great potential to improve the performance of, and have fewer tuning parameters than, the directional TV term used in this study.

Another good alternative is to use the Fourier structure of the sampling pattern more heavily in the design of a regulariser. Bubba et al. (2019) pursue this approach where the authors use a Shearlet frame to separate the coefficients corresponding to the denoising/inpainting components of data and then treat each partition independently. The results showed great promise but never attempted such a severely ill-posed regime, at most 45 % missing data as opposed to 66 % in our experiments.

Shearlets provide the perfect means to distinguish the denoising and inpainting problems, although it remains unclear what a good inpainting model would be. The solution of Bubba et al. (2019) was to learn such a model. In applications, the aim is to extract the information that is present in the data as clearly as possible. The missing data cannot be known but a good reconstruction should not allow the missing data to corrupt the interpretation of the observed data. Within this remit, I am hopeful that a classical $L^1$-sparsity model exists which accurately reproduces the desired features whilst always localising the damage from the missing wedge.

### 7.1.2 Optimisation algorithm

In the current era of machine learning, one of the most common issues raised against TV reconstructions is the long computational time. In this context, the numerical scheme proposed requires the solution of the order of 100 TV reconstructions. The cause of this is the iterative estimation of the directional component, each iteration requires the solution of two TV-like optimisation problems. The biggest improvement to this would be to perform a single iterative scheme which refines the reconstruction and structure tensor more gradually but much more cheaply.

If the formulation that we proposed is intrinsically very slow to optimise, then it is possible that another formulation, perhaps using a modification of the regulariser discussed above, is more amenable to numerical minimisation. There is a definite interest in the community to have a reconstruction method which is stable with a large missing wedge and there will always be space for such a method which takes a long time to run. On the other hand, as argued before, one imagines that there should be a good but imperfect convex method which runs quickly with good localisation of errors. Stochastic optimisation methods, such as those proposed by Chambolle et al. (2018) in the convex setting, may be one direction to explore.

## 7.2 Strain tomography

In Chapter 3 we proposed a new physical model for the reconstruction of strain maps from electron diffraction data. The forward model was validated numerically with simulated data and the inverse problem was also shown to be accurately solvable.

### 7.2.1 Forward model

The validation accuracy of the forward modelling is very promising. The practical implication is that the simple linearised transverse ray transform approximation is at least as accurate as the diffraction simulation model we used.

The largest drop in accuracy occurred when the distribution of strain became smooth instead of piecewise constant. Visually, the strain blurred out circular spots into ellipses which could not happen with a discrete distribution of strain (Figure B.2 vs. Figure 3.3). Whether or not this is the cause, such strains are common in practice so the behaviour needs to be better understood. This does not occur in 2D-like structures which is possibly why the phenomenon does not appear to be well-explored to this point.

In terms of experimental practicality, the greatest limitation is guaranteeing that each tilt aligns exactly with a zone-axis. It is time consuming to tune the microscope so finely, so many times. This assumption is essential for the correspondence with the transverse ray transform, but not anything more fundamental. The transverse ray transform is convenient because there is only one direction of interest; the beam direction. Data is integrated along the beam and is insensitive to strain orthogonal to the beam. In *off-zone-axis* diffraction patterns, the data will still be integrated along the beam but there will be a new direction dictating the insensitivity. The centre of mass model will still be linear but the map will no longer coincide with the transverse ray transform. It would be interesting to know whether the properties of the new linearisation would be any better or worse than the one studied in Section 3.7.

### 7.2.2 Reconstruction

The mathematical theory of tomography typically centres on continuous tilt series with an infinite amount of data. In the case of diffraction imaging the tilt series is intrinsically discrete, there is little insight as to the behaviour of such a scan. It would be nice to at least know asymptotic bounds to see how error scales with number of tilts. Alternatively, if the samples look ‘sufficiently random’, can the theory of compressed sensing be used? For now it is sufficient to understand such sampling schemes from a numerical standing, however, mathematical guarantees would represent a great step forward.

Other than the lack of theoretical insight, the numerical results were convincing and indicate that the inverse problem can be solved effectively in practice.

## 7.3 Adaptive FISTA

In Chapter 4 we propose a new optimisation algorithm for computing Banach space (or high resolution) minimisers. This enables FISTA to be extended to a new class of problems with relatively sharp convergence guarantees. The adaptive scheme is consistently faster and more efficient with computational resources than using a classical high resolution discretisation.

### 7.3.1 Amenable forward models

While non-uniform discretisations have the potential to be more efficient, this also relies on code being capable of utilising that structure. In the case of Lasso, and in most inverse problems, the challenge will be the forward map. The cost of naive matrix-vector multiplication scales with the product of data size and reconstruction dimension; this grows very quickly for large data and complex reconstructions. In large-scale applications, one relies on highly efficient or 'fast' variants of code. For instance, there are very high performing packages for tomography but they are designed for uniform pixel size. Alternatively, there is the fast Fourier transform which has been revolutionary in imaging applications, but is very difficult to implement when the domain is constantly adapting.

In this work we used Gaussian convolution and the X-ray transform which both work well because the underlying operator is sparse and the complexity scaling can be well-controlled. The Fourier kernel implementation is much more naive and only efficient when the number of data points is small.

Unfortunately, the image processing community has remained quite distinct from the finite element community where it is common to require flexible discretisations. While there are many individuals who use finite elements in imaging, for example Arridge et al. (1993); Bartels (2012); Carrascal-Manzanares et al. (2018); Monard et al. (2019), the imaging community has not followed the PDE community in their mainstream adoption[1][2][3]. Until there exist efficient toolboxes which allow for the implementation of adaptive methods as easily as current standard methods, adaptively discretised schemes are unlikely to achieve widespread usage.

### 7.3.2 Further Lasso specialisation

While the adaptive FISTA algorithm is not limited to the continuous Lasso problem, this is an example where the concept of infinite resolution image processing is well studied analytically

[1] freefem.org/

[2] fenicsproject.org/

[3] www.firedrakeproject.org/

and numerically. The standard approach is to discretise with a sum of Dirac deltas, which can be exact with a finite number of parameters. Optimisation is then performed by solving a sequence of non-convex optimisation problems, adding one new delta at each iteration.

Such a discrete basis is clearly very specific and memory efficient for the Lasso problem, although it is hard to derive rates or even global convergence guarantees. Until the whole support of $u^*$ is identified, there cannot be any convergence. In my opinion, there are also insufficient posteriori checks performed to guarantee if the support is identified. The necessary check is to plot the gradient of the data term and guarantee it is no larger than $\mu$ at any point, however this verification is not seen in many studies (Bredies and Pikkarainen, 2013; De Castro et al., 2016; Boyd et al., 2017; Huang et al., 2017; Catala et al., 2019; Denoyelle et al., 2019).

The way we account for this problem in Chapter 4 is with a Taylor expansion performed on every pixel of the mesh. The Taylor expansion provides upper-bound error estimates which is used to guarantee the appropriate global convergence rate. If the error bound were combined with some geometrical properties of Lasso (Candès and Fernandez-Granda, 2014; Poon et al., 2018), then this would identify when every spike of the support had been isolated.

If the discretisations could be merged, for instance a refining mesh with one Dirac delta per pixel, I think the two approaches could complement each other greatly. The Dirac basis allows for a very direct link between the discrete and continuum problems, while the mesh allows for support identification and preventing premature overfitting to the discrete problem. Some of these deltas will be driven to identify the support, as before, while the excess find the best Taylor expansion points for the purpose of excluding pixels from the support.

### 7.3.3 Other algorithms

Once the concept of an appropriate 'back-stop' became solidified, remaining proofs followed very quickly. The arguments to prove rate estimates combined the classical rate estimations of FISTA with an inductive step to bound error growth, thus determining the scaling of the back-stop. This framework should generalise easily to other optimisation algorithms such as primal-dual (Chambolle and Pock, 2011) and Douglas-Rachford (Combettes and Pesquet, 2011).

Many practical modifications like restarting and line-search should also be possible. If the classical method has a convergence rate then a refinement strategy can be chosen to achieve a slightly smaller rate.

### 7.3.4 Link to stochastic optimisation

The mantra of stochastic optimisation is that each iteration of an algorithm is allowed to be inexact, so long as the inexactness is does not correlate with the $u^*$ (i.e. the estimate is 'un-biased'). If this is true then the convergence rate should be the same as the original algorithm but only in a slightly weaker sense (an 'on average' sense).

The corresponding mantra of this work is that the inexactness is only allowed to be a blurring of the exact case. The cost for this in the Hilbert-space case is that the errors must decay sufficiently quickly to achieve the correct rate, again in a slightly weaker sense.

I am intrigued by the question of whether there is any way to interpret the coarse approximation as an un-biased approximation. If the discretisation is always uniform and refined 'quickly enough', then is there any bias between the exact or inexact output after $n$ iterations? My difficulty with this is that there is no intrinsic randomness to the iterations. Maybe considered over the distribution of functionals such that $\Pi_n u^*$ is fixed?

## 7.4 Total variation discretisation

In Chapter 5 we introduced a new benchmark example which highlights the limitation of current state-of-the-art methods, and we proposed a new parametrisation with the potential to overcome the newly identified limitation. Numerical results comparing the new example with previous experiments are very clear and sharply achieve the known lower-bound convergence rate. The new finite element is not practical and still very much a work in progress, but its existence is interesting in its own right. It provides a convex formulation for examples of adaptive-mesh finite element methods and piecewise linear estimation of level sets. Both tasks are typically non-convex with many bad non-local optima, I do not know how unique it is to have found a convex formulation.

### 7.4.1 Benchmark examples

The extension of the new example to higher dimensions is trivial, replacing circles with spheres. The remaining question is whether this example is also enough to achieve the lower bound rate of Bartels (2015) (assuming $L^\infty$ data) in dimensions greater than 2. The generic argument of Section 5.3.2 suggests that this should be the case. Beyond confirming this detail, I don't think there are any more analytical gains to be made in this area.

### 7.4.2 TV-optimal finite element

There is much more work to be done on this problem, although I believe there is a place for it in the image processing community. Total variation is often used in imaging to identify when edges exist and where they are. In some sense, the proposed discretisation should over-accentuate these features by producing sharp interfaces at coarse resolutions. Linking with Chapter 4, an improved convergence rate means that a coarser resolution can be used to compute equally accurate results. If the complexity of the discretisation is not much more complex than a standard piecewise linear finite element method, then this should correspond to faster computation times as well.

## 7.5 Reconstruction with a Gaussian dictionary

In Chapter 6 we performed some numerical comparisons on the inverse problem of atomic reconstructions in single particle analysis. The initial results demonstrate interesting trends but require further verification.

### 7.5.1 Optimal scheme

The results suggested that the optimal scheme would be to start by fitting Gaussians directly to the data to locate the majority of atoms, then use this as an initialisation to fitting atoms to the volume reconstruction. Gaussians should be assigned large fixed radii and only optimised for location and total charge. The first optimisation should be performed with Newton then the second possibly with a stochastic method to find the remaining atoms.

This is a perfectly feasible strategy although I am not sure how it will compare with other more customised methods. If the final optimisation problem is just a three dimensional Gaussian mixture decomposition, then this is a well studied problem with many good initialisation strategies. It seems unlikely that combining the initialisation process with the X-ray transform will lead to an improved output after the refinement process.

### 7.5.2 Potential of direct fitting

At the beginning of this work it was anticipated that directly fitting atomic reconstructions to raw data must only result in better accuracy than fitting to an intermediate reconstruction. In general this seems to be false. This is mainly due to the non-convexity of the problem combined with the properties of the X-ray transform. Fitting Gaussian centres through the X-ray transform means that the location of each Gaussian is biased by all errors throughout the volume, including parts of the volume which are poorly reconstructed. This is hard to avoid in non-convex optimisation yet limits the accuracy of the local minimiser much more strongly than in sequential reconstructions.

It may not be possible to avoid the locality issue with direct fitting. Alternatively, this can be restated as: the prior that atoms look like Gaussians is strong enough to overcome the limitations of fitting to an intermediate reconstruction. Aside from the poor robustness in our basic implementations, sequential fitting is capable of very accurate reconstructions from a very primitive least squares volume reconstruction. In retrospect, this is probably not surprising due to the very large quantity of data. The angular resolution is very high which should minimise the missing wedge and super-resolution issues typically seen in other applications. The regularising effect of reconstructing in a Gaussian basis is clearly capable of accounting for the remaining artifacts and noise.

## 7.6 Concluding remarks

My work over the last four years has been in contribution to the better understanding and solving of problems relating to electron tomography.

I think that the proposed limited angle model represents a useful addition to previous methods, although it is a topic which deserves revisiting due to its importance, not just in electron microscopy but also across other modalities. It is a problem that cannot be avoided with hardware and deserves a simple and reliable baseline reconstruction method.

The new diffraction model arises in a background where I think the community is beginning to realise that the historical models for EM are not good enough, and the hardware is capable of doing more. Cutting edge diffraction imaging often felt to me like a heuristic analysis; the fundamentals are solid and very mathematically advanced, but there is also a large gap between theory and practice. Understanding diffraction from a tomographic perspective required a narrow bridge to be built over this gap but I hope there is more to come. 50 years on from its first introduction to EM, the X-ray transform is also in dire need for an update. Data is currently discarded if it doesn't 'look linear enough', but these are the images which contain most information and it is a very wasteful practice. Some modifications can still result in a convex optimisation reconstruction framework, but it is also possible to go beyond this with modern computing power.

I will take this as an opportunity to touch upon a couple of other aspects of electron microscopy which are currently of much interest to the community but which could not be investigated during my PhD. Microscopists typically record many datasets of the same object, i.e. *multi-modal* data, but combine the information manually rather than allowing each dataset to improve the reconstruction of the others. This idea has already been implemented successfully in many areas outside of EM (Ehrhardt and Arridge, 2013; Ehrhardt et al., 2015; Merlin et al., 2018). The EM community has begun to explore this direction (Guo et al., 2019; Starborg et al., 2019), but there is not yet sufficient communication with communities where this problem has already received much more attention. Another theme that has become apparent to me is the desire to ask statistical questions of data, not just to compute a single reconstruction. A good example is where a sample may consist of two spherical object and the question is whether they touch or not. A reconstruction will just show a most likely outcome, but not attach a certainty to it. The root of this question is hidden in the sensitivity of that (possible) touching point in the reconstruction to the data, but it is very difficult to interpret from a single image. The standard approach for this is switch to a statistical model, as opposed to optimisation, and perform what is known as sampling from the posterior-distribution (Latz et al., 2018). In essence, computing many reconstructions which are all physically likely and fit the data. This process allows accurate assignment of probabilities to events although, in many applications, it is prohibitively expensive. Regardless of the expense, there is definitely interest in knowing how to address these questions in electron microscopy.

Optimisation will always be an important component of inverse problems, and therefore also electron microscopy, so long as people desire the ‘best’ reconstruction. Embedded within this is the three-way arms race to create more powerful models, more powerful computers, and more efficient algorithms. I think that the engineers are currently losing this battle due to the large quantities of data in every potential application. I wish that more models were designed without the need for manual parameter tuning, however, in a field where TV reconstruction is still state of the art, the limitation is the power of computation rather than available models. This was the motivation for investigating an adaptive discretisation in imaging. I believe that the PDE community is around 20-30 years ahead of the imaging community in the realisation that pixels do not have to be square or flat. That is not to say that no one in the imaging community pursues this idea, but there are no wide-spread packages available which enable non-specialists to benefit. On the other hand, the pressure of ‘big-data’ and competition from machine learning means that efficiency is more important than ever. The results I have seen during my PhD make me very hopeful for the potential of this type of adaptive algorithms and it is something I wish to pursue further in the future.

# References

Abhishek, A. (2020). Support theorems for the transverse ray transform of tensor fields of rank m. *Journal of Mathematical Analysis and Applications*, 485(2):123828.

Agulleiro, J. I., Garzón, E. M., García, I., and Fernández, J. J. (2010). Vectorization with SIMD extensions speeds up reconstruction in electron tomography. *Journal of Structural Biology*, 170(3):570–575.

Ambrosio, L., Caselles, V., Masnou, S., and Morel, J.-M. (2001). Connected components of sets of finite perimeter and applications to image processing. *Journal of the European Mathematical Society*, 3(1):39–92.

Ambrosio, L., Fusco, N., and Pallara, D. (2000). *Functions of Bounded Variation and Free Discontinuity Problems*, volume 254. Clarendon Press Oxford.

Arridge, S., Maass, P., Öktem, O., and Schönlieb, C.-B. (2019). Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174.

Arridge, S., Schweiger, M., Hiraoka, M., and Delpy, D. (1993). A finite element approach for modeling photon transport in tissue. *Medical Physics*, 20(2):299–309.

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. (2010). Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457.

Aujol, J.-F. and Dossal, C. (2015). Stability of over-relaxations for the forward-backward algorithm, application to fista. *SIAM Journal on Optimization*, 25(4):2408–2433.

Bartels, S. (2012). Total variation minimization with finite elements: Convergence and iterative solution. *SIAM Journal on Numerical Analysis*, 50(3):1162–1180.

Bartels, S. (2015). *Numerical Methods for Nonlinear Partial Differential Equations*, volume 47. Springer.

Bartels, S. (2020). Nonconforming discretizations of convex minimization problems and precise relations to mixed methods. *arXiv preprint arXiv:2002.02359*.

Béché, A., Rouvière, J. L., Barnes, J. P., and Cooper, D. (2013). Strain measurement at the nanoscale: Comparison between convergent beam electron diffraction, nano-beam electron diffraction, high resolution imaging and dark field electron holography. *Ultramicroscopy*, 131:10–23.

Béché, A., Rouvière, J. L., Clément, L., and Hartmann, J. M. (2009). Improved precision in strain measurement using nanobeam electron diffraction. *Applied Physics Letters*, 95(12):123114.

Beck, A. and Teboulle, M. (2009). Gradient-based algorithms with applications to signal recovery problems. In *Convex Optimization in Signal Processing and Communications*, pages 42–88. Cambridge University Press.

Benning, M. and Burger, M. (2018). Modern regularization methods for inverse problems. *Acta Numerica*, pages 1–111.

Berkels, B., Burger, M., Droske, M., Nemitz, O., and Rumpf, M. (2006). Cartoon Extraction Based on Anisotropic Image Classification. *Vision, Modeling, and Visualization*, page 293.

Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C. (2000). Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '00*, pages 417–424. ACM Press.

Bolte, J., Daniilidis, A., Lewis, A., and Shiota, M. (2007). Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572.

Bolte, J., Sabach, S., and Teboulle, M. (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494.

Boman, J. and Sharafutdinov, V. (2018). Stability estimates in tensor tomography. *Inverse Problems & Imaging*, 12(5):1245.

Bonef, B., Haas, B., Rouvière, J.-L., André, R., Bougerol, C., Grenier, A., Jouneau, P.-H., and Zuo, J.-M. (2016). Interfacial chemistry in a ZnTe/CdSe superlattice studied by atom probe tomography and transmission electron microscopy strain measurements. *Journal of Microscopy*, 262(2):178–182.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311.

Boyd, N., Schiebinger, G., and Recht, B. (2017). The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616–639.

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Boyer, C., Chambolle, A., Castro, Y. D., Duval, V., De Gournay, F., and Weiss, P. (2019). On representer theorems and convex regularization. *SIAM Journal on Optimization*, 29(2):1260–1281.

Bracewell, R. N. (1956). Strip integration in radio astronomy. *Australian Journal of Physics*, 9(2):198–217.

Bredies, K. and Pikkarainen, H. K. (2013). Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218.

Bubba, T. A., Kutyniok, G., Lassas, M., März, M., Samek, W., Siltanen, S., and Srinivasan, V. (2019). Learning the invisible: a hybrid deep learning-shearlet framework for limited angle computed tomography. *Inverse Problems*, 35(6):064002.

Burger, M., Müller, J., Papoutsellis, E., and Schönlieb, C.-B. (2014). Total variation regularisation in measurement and image space for PET reconstruction. *Inverse Problems*, 30(10).

Calatroni, L., Lanza, A., Pragliola, M., and Sgallari, F. (2019). A flexible space-variant anisotropic regularization for image restoration with automated parameter selection. *SIAM Journal on Imaging Sciences*, 12(2):1001–1037.

Candes, E., Demanet, L., Donoho, D., and Ying, L. (2006). Fast discrete curvelet transforms. *Multiscale Modeling & Simulation*, 5(3):861–899.

Candès, E. J. and Donoho, D. L. (2004). New tight frames of curvelets and optimal representations of objects with piecewise $c^2$ singularities. *Communications on Pure and Applied Mathematics*, 57(2):219–266.

Candès, E. J. and Fernandez-Granda, C. (2014). Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956.

Candès, E. J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509.

Carrascal-Manzanares, C., Imperiale, A., Rougeron, G., Bergeaud, V., and Lacassagne, L. (2018). A fast implementation of a spectral finite elements method on cpu and gpu applied to ultrasound propagation. *Advances in Parallel Computing*, pages 339–348.

Castillo, I. and Rockova, V. (2019). Multiscale analysis of bayesian cart. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2019-127).

Catala, P., Duval, V., and Peyré, G. (2019). A low-rank approach to off-the-grid sparse superresolution. *SIAM Journal on Imaging Sciences*, 12(3):1464–1500.

Chambolle, A. and Dossal, C. (2015). On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166(3):968–982.

Chambolle, A., Duval, V., Peyré, G., and Poon, C. (2016). Geometric properties of solutions to the total variation denoising problem. *Inverse Problems*, 33(1):015002.

Chambolle, A., Ehrhardt, M. J., Richtárik, P., and Schonlieb, C.-B. (2018). Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808.

Chambolle, A., Levine, S. E., and Lucier, B. J. (2009). Some variations on total variation-based image smoothing. Technical report, Minnesota univ. Minneapolis Inst. for Mathematics and its Applications.

Chambolle, A. and Lions, P.-L. (1997). Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188.

Chambolle, A. and Pock, T. (2011). A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145.

Chambolle, A. and Pock, T. (2020). Crouzeix–raviart approximation of the total variation on simplicial meshes. *Journal of Mathematical Imaging and Vision*, pages 1–28.

Chan, R. H., Chan, T. F., Shen, L., and Shen, Z. (2003). Wavelet algorithms for high-resolution image reconstruction. *SIAM Journal on Scientific Computing*, 24(4):1408–1432.

Chan, T. F., Kang, S. H., and Shen, J. (2001). Total variation denoising and enhancement of color images based on the cb and hsv color models. *Journal of Visual Communication and Image Representation*, 12(4):422–435.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335.

Chu, M., Sun, Y., Aghoram, U., and Thompson, S. E. (2009). Strain: A Solution for Higher Carrier Mobility in Nanoscale MOSFETs. *Annual Review of Materials Research*, 39(1):203–229.

Claerbout, J. F. and Muir, F. (1973). Robust modeling with erratic data. *Geophysics*, 38(5):826–844.

Clarke, F. H. (1990). *Optimization and Nonsmooth Analysis*. SIAM.

Cnudde, V. and Boone, M. N. (2013). High-resolution x-ray computed tomography in geosciences: A review of the current technology and applications. *Earth-Science Reviews*, 123:1–17.

Collins, S. M., Leary, R. K., Midgley, P. A., Tovey, R., Benning, M., Schönlieb, C.-B., Rez, P., and Treacy, M. M. (2017). Entropic Comparison of Atomic-Resolution Electron Tomography of Crystals and Amorphous Materials. *Physical Review Letters*, 119(16):166101.

Collins, S. M., MacArthur, K. E., Longley, L., Tovey, R., Benning, M., Schönlieb, C.-B., Bennett, T. D., and Midgley, P. A. (2019). Phase diagrams of liquid-phase mixing in multi-component metal-organic framework glasses constructed by quantitative elemental nano-tomography. *APL Materials*, 7(9):091111.

Collins, S. M., Ringe, E., Duchamp, M., Saghi, Z., Dunin-Borkowski, R. E., and Midgley, P. A. (2015). Eigenmode tomography of surface charge oscillations of plasmonic nanoparticles by electron energy loss spectroscopy. *ACS Photonics*, 2(11):1628–1635.

Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer.

Committee (1992). Report of the Executive Committee for 1991. *Acta Crystallographica*, 48(6):922–946.

Condat, L. (2017). Discrete total variation: New definition and minimization. *SIAM Journal on Imaging Sciences*, 10(3):1258–1290.

Cooper, D., Bernier, N., and Rouvière, J.-L. (2015). Combining 2 nm Spatial Resolution and 0.02% Precision for Deformation Mapping of Semiconductor Specimens in a Transmission Electron Microscope by Precession Electron Diffraction. *Nano Letters*, 15(8):5289–5294.

Cooper, D., Bernier, N., Rouvière, J.-L., Wang, Y.-Y., Weng, W., Madan, A., Mochizuki, S., and Jagannathan, H. (2017). High-precision deformation mapping in finFET transistors with two nanometre spatial resolution by precession electron diffraction. *Applied Physics Letters*, 110(22):223109.

Cowley, J. M. (1981). *Diffraction Physics*. North-Holland, New York, NY, second edition.

Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. (2007). Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095.

Dahlin, J., Lindsten, F., and Schön, T. B. (2015). Particle metropolis–hastings using gradient and hessian information. *Statistics and Computing*, 25(1):81–92.

De Castro, Y., Gamboa, F., Henrion, D., and Lasserre, J.-B. (2016). Exact solutions to super resolution on semi-algebraic domains in higher dimensions. *IEEE Transactions on Information Theory*, 63(1):621–630.

De Rosier, D. and Klug, A. (1968). Reconstruction of three dimensional structures from electron micrographs. *Nature*, 217(5124):130–134.

Deans, S. R. (1983). *The Radon Transform and Some of Its Applications.* Wiley.

Demers, H., Poirier-Demers, N., Couture, A. R., Joly, D., Guilmain, M., de Jonge, N., and Drouin, D. (2011). Three-dimensional electron microscopy simulation with the casino monte carlo software. *Scanning*, 33(3):135–146.

Denoyelle, Q., Duval, V., Peyré, G., and Soubies, E. (2019). The sliding frank–wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001.

Desai, N. M. and Lionheart, W. R. B. (2016). An explicit reconstruction algorithm for the transverse ray transform of a second rank tensor field from three axis data. *Inverse Problems*, 32(11):115009.

Donati, L., Nilchian, M., Sorzano, C. O. S., and Unser, M. (2018). Fast multiscale reconstruction for cryo-em. *Journal of structural biology*, 204(3):543–554.

Dong, B., Li, J., and Shen, Z. (2013). X-Ray CT Image Reconstruction via Wavelet Frame Based Regularization and Radon Domain Inpainting. *Journal of Scientific Computing*, 54(2-3):333–349.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306.

Drusvyatskiy, D., Ioffe, A. D., and Lewis, A. S. (2019). Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria. *Mathematical Programming*, pages 1–27.

Drusvyatskiy, D. and Lewis, A. S. (2018). Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948.

Duchi, J. C. (2017). Introductory Lectures on Stochastic Population Systems. *arXiv preprint arXiv:1705.03781*, pages 1–84.

Dunbar, O. R., Dunlop, M. M., Elliott, C. M., Hoang, V. H., and Stuart, A. M. (2020). Reconciling bayesian and perimeter regularization for binary inversion. *SIAM Journal on Scientific Computing*, 42(4):A1984–A2013.

Duval, V. and Peyré, G. (2017a). Sparse spikes super-resolution on thin grids i: the lasso. *Inverse Problems*, 33(5):055008.

Duval, V. and Peyré, G. (2017b). Sparse spikes super-resolution on thin grids ii: the continuous basis pursuit. *Inverse Problems*, 33(9):095008.

Egerton, R. F. (2005). *Physical Principles of Electron Microscopy*, volume 56. Springer.

Ehrhardt, M. J. and Arridge, S. R. (2013). Vector-valued image processing by parallel level sets. *IEEE Transactions on Image Processing*, 23(1):9–18.

Ehrhardt, M. J., Thielemans, K., Pizarro, L., Atkinson, D., Ourselin, S., Hutton, B. F., and Arridge, S. R. (2015). Joint reconstruction of PET-MRI by exploiting structural similarity. *Inverse Problems*, 31(1):015001.

Ekanadham, C., Tranchina, D., and Simoncelli, E. P. (2011). Recovery of sparse translation-invariant signals with continuous basis pursuit. *IEEE transactions on signal processing*, 59(10):4735–4744.

Engl, H. W. and Grever, W. (1994). Using the l–curve for determining optimal regularization parameters. *Numerische Mathematik*, 69(1):25–31.

Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems*, volume 375. Springer Science & Business Media.

Estellers, V., Soatto, S., and Bresson, X. (2015). Adaptive Regularization With the Structure Tensor. *IEEE Transactions on Image Processing*, 24(6):1777–1790.

Fokas, A., Iserles, A., and Marinakis, V. (2005). Reconstruction algorithm for single photon emission computed tomography and its numerical implementation. *Journal of the Royal Society Interface*.

Freeman, W. T., Jones, T. R., and Pasztor, E. C. (2002). Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65.

Freundlich, M. M. (1963). Origin of the electron microscope. *Science*, 142(3589):185–188.

Frikel, J. and Quinto, E. T. (2013). Characterization and reduction of artifacts in limited angle tomography. *Inverse Problems*, 29(12):125007.

Galindo, P. L., Kret, S., Sanchez, A. M., Laval, J.-Y., Yanez, A., Pizarro, J., Guerrero, E., Ben, T., and Molina, S. I. (2007). The peak pairs algorithm for strain mapping from hrtem images. *Ultramicroscopy*, 107(12):1186–1193.

Gilbert, P. (1972). Iterative methods for the three-dimensional reconstruction of an object from projections. *Journal of Theoretical Biology*, 36(1):105–117.

Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons.

Giustino, F. (2014). *Materials Modelling using Density Functional Theory: Properties and Predictions*. Oxford University Press.

Goris, B., De Beenhouwer, J., De Backer, A., Zanaga, D., Batenburg, K. J., Sánchez-Iglesias, A., Liz-Marzán, L. M., Van Aert, S., Bals, S., Sijbers, J., and Van Tendeloo, G. (2015). Measuring lattice strain in three dimensions through electron microscopy. *Nano Letters*, 15(10):6996–7001.

Goris, B., Van den Broek, W., Batenburg, K. J., Heidari Mezerji, H., and Bals, S. (2012). Electron tomography based on a total variation minimization reconstruction technique. *Ultramicroscopy*, 113:120–130.

Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In Blondel, V., Boyd, S., and Kimura, H., editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited.

Grant, M. and Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx.

Gu, J. and Ye, J. C. (2017). Multi-Scale Wavelet Domain Residual Learning for Limited-Angle CT Reconstruction. *arXiv preprint arXiv:1703.01382*.

Gu, J., Zhang, L., Yu, G., Xing, Y., and Chen, Z. (2006). X-ray CT metal artifacts reduction through curvature based sinogram inpainting. *Journal of X-ray Science and Technology*, 14(2):73–82.

Guo, K. and Labate, D. (2007). Optimally sparse multidimensional representation using shearlets. *SIAM Journal on Mathematical Analysis*, 39(1):298–318.

Guo, Y., Aveyard, R., and Rieger, B. (2019). A multichannel cross-modal fusion framework for electron tomography. *IEEE Transactions on Image Processing*, 28(9):4206–4218.

Haas, B., Thomas, C., Jouneau, P.-H., Bernier, N., Meunier, T., Ballet, P., and Rouvière, J.-L. (2017). High precision strain mapping of topological insulator HgTe/CdTe. *Applied Physics Letters*, 110(26):263102.

Hammernik, K., Würfl, T., Pock, T., and Maier, A. (2017). A Deep Learning Architecture for Limited-Angle Computed Tomography Reconstruction. In *Bildverarbeitung für die Medizin 2017*, pages 92–97. Springer Vieweg, Berlin, Heidelberg.

Helgason, S. (1980). *The radon transform*, volume 2. Springer.

Hintermüller, M., Holler, M., and Papafitsoros, K. (2018). A function space framework for structural total variation regularization with applications in inverse problems. *Inverse Problems*, 34(6):064002.

Hintermüller, M., Papafitsoros, K., and Rautenberg, C. N. (2017). Analytical aspects of spatially adapted total variation regularisation. *Journal of Mathematical Analysis and Applications*, 454(2):891–935.

Hirsch, P. B., Howie, A., Nicholson, R. B., Pashley, D. W., and Whelan, M. J. (1967). *Electron microscopy of thin crystals*. Butterworths, London.

Högbom, J. (1974). Aperture synthesis with a non-regular distribution of interferometer baselines. *Astronomy and Astrophysics Supplement Series*, 15:417.

Holman, S. (2013). Generic local uniqueness and stability in polarization tomography. *Journal of Geometric Analysis*, 23(1):229–269.

Hou, H. and Andrews, H. (1978). Cubic splines for image interpolation and digital filtering. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(6):508–517.

Howie, A. and Basinski, Z. S. (1968). Approximations of the dynamical theory of diffraction contrast. *The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics*, 17(149):1039–1063.

Hsieh, D., Xia, Y., Qian, D., Wray, L., Dil, J. H., Meier, F., Osterwalder, J., Patthey, L., Checkelsky, J. G., Ong, N. P., Fedorov, A. V., Lin, H., Bansil, A., Grauer, D., Hor, Y. S., Cava, R. J., and Hasan, M. Z. (2009). A tunable topological insulator in the spin helical Dirac transport regime. *Nature*, 460(7259):1101–1105.

Hÿtch, M. J. and Minor, A. M. (2014). Observing and measuring strain in nanostructures and devices with transmission electron microscopy. *MRS Bulletin; Warrendale*, 39(2).

Huang, J., Sun, M., Ma, J., and Chi, Y. (2017). Super-resolution image reconstruction for high-density three-dimensional single-molecule microscopy. *IEEE Transactions on Computational Imaging*, 3(4):763–773.

Hÿtch, M., Houdellier, F., Hüe, F., and Snoeck, E. (2011). Dark-field electron holography for the measurement of geometric phase. *Ultramicroscopy*, 111(8):1328–1337.

Hÿtch, M. J., Putaux, J.-L., and Pénisson, J.-M. (2003). Measurement of the displacement field of dislocations to 0.03 å by electron microscopy. *Nature*, 423(6937):270–273.

Jiang, K., Sun, D., and Toh, K.-C. (2012). An inexact accelerated proximal gradient method for large scale linearly constrained convex sdp. *SIAM Journal on Optimization*, 22(3):1042–1064.

John, F. (1938). The ultrahyperbolic differential equation with four independent variables. *Duke Math. J.*, 4(2):300–322.

Johnstone, D. N., van Helvoort, A. T. J., and Midgley, P. A. (2017). Nanoscale Strain Tomography by Scanning Precession Electron Diffraction. *Microscopy and Microanalysis*, 23(S1):1710–1711.

Joubert, P. and Habeck, M. (2015). Bayesian inference of initial models in cryo-electron microscopy using pseudo-atoms. *Biophysical Journal*, 108(5):1165–1175.

Kaipio, J. P., Kolehmainen, V., Vauhkonen, M., and Somersalo, E. (1999). Inverse problems with structural prior information. *Inverse Problems*, 15(3):713–729.

Kalender, W. A. (2006). X-ray computed tomography. *Physics in Medicine & Biology*, 51(13).

Kalke, M. and Siltanen, S. (2014). Sinogram Interpolation Method for Sparse-Angle Tomography. *Applied Mathematics*, 05(03):423–441.

Katsevich, A. I. (1997). Local tomography for the limited-angle problem. *Journal of Mathematical Analysis and Applications*, 213(1):160–182.

Kim, L. Y., Rice, W. J., Eng, E. T., Kopylov, M., Cheng, A., Raczkowski, A. M., Jordan, K. D., Bobe, D., Potter, C. S., and Carragher, B. (2018). Benchmarking cryo-em single particle analysis workflow. *Frontiers in Molecular Biosciences*, 5:50.

Kirkland, E. J. (1998). *Advanced Computing in Electron Microscopy.* Springer.

Kohn, W. and Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133.

Korsunsky, A. M., Baimpas, N., Song, X., Belnoue, J., Hofmann, F., Abbey, B., Xie, M., Andrieux, J., Buslaps, T., and Neo, T. K. (2011). Strain tomography of polycrystalline zirconia dental prostheses by synchrotron x-ray diffraction. *Acta Materialia*, 59(6):2501–2513.

Korsunsky, A. M., Vorster, W. J., Zhang, S. Y., Dini, D., Latham, D., Golshan, M., Liu, J., Kyriakoglou, Y., and Walsh, M. J. (2006). The principle of strain reconstruction tomography: Determination of quench strain distribution from diffraction measurements. *Acta Materialia*, 54(8):2101–2108.

Köstler, H., Prümmer, M., Rüde, U., and Hornegger, J. (2006). Adaptive variational sinogram interpolation of sparsely sampled CT data. In *Proceedings - International Conference on Pattern Recognition*, volume 3, pages 778–781. IEEE.

Krishnan, V. P. and Quinto, E. T. (2015). Microlocal analysis in tomography. *Handbook of Mathematical Methods in Imaging*, 1:2.

Kübel, C., Voigt, A., Schoenmakers, R., Otten, M., Su, D., Lee, T.-C., Carlsson, A., and Bradley, J. (2005). Recent advances in electron tomography: TEM and HAADF-STEM tomography for materials science and semiconductor applications. *Microscopy and microanalysis*, 11(5):378–400.

Kutyniok, G. and Labate, D. (2012). *Shearlets: Multiscale Analysis for Multivariate Data*. Springer Science & Business Media.

Larson, B., Yang, W., Ice, G., Budai, J., and Tischler, J. (2002). Three-dimensional x-ray structural microscopy with submicrometre resolution. *Nature*, 415(6874):887–890.

Latz, J., Papaioannou, I., and Ullmann, E. (2018). Multilevel sequential$^2$ monte carlo for bayesian inverse problems. *Journal of Computational Physics*, 368:154–178.

Leary, R. K. and Midgley, P. A. (2019). *Electron Tomography in Materials Science*, pages 2–2. Springer International Publishing.

Leary, R. K., Saghi, Z., Midgley, P. A., and Holland, D. J. (2013). Compressed sensing electron tomography. *Ultramicroscopy*, 131:70–91.

Lebrun, M., Colom, M., Buades, A., and Morel, J.-M. (2012). Secrets of image denoising cuisine. *Acta Numerica*, 21:475.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690.

Levine, Z. H. (2005). Theory of bright-field scanning transmission electron microscopy for tomography. *Journal of Applied Physics*, 97(3):033101.

Lewis, G. R., Loudon, J., Tovey, R., Chen, Y.-H., Harrison, R., Midgley, P., and Ringe, E. (2020). Magnetic vortex states in toroidal iron oxide nanoparticles: Combining micromagnetics with tomography. accepted in Nano letters 2020.

Li, S., Cao, Q., Chen, Y., Hu, Y., Luo, L., and Toumoulin, C. (2014). Dictionary learning based sinogram inpainting for ct sparse reconstruction. *Optik*, 125(12):2862–2867.

Liang, J., Fadili, J. M., and Peyré, G. (2016). A Multi-step Inertial Forward-Backward Splitting Method for Non-convex Optimization. *Advances in Neural Information Processing Systems*.

Liang, J. and Schönlieb, C.-B. (2018). Improving fista: Faster, smarter and greedier. *arXiv preprint arXiv:1811.01430*.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., et al. (2019). Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in phenix. *Acta Crystallographica Section D: Structural Biology*, 75(10):861–877.

Lionheart, W. R. (2019). Histogram tomography. *Mathematics in Engineering*, 2:55–74.

Lionheart, W. R. B. and Withers, P. J. (2015). Diffraction tomography of strain. *Inverse Problems*, 31(4):045005.

Lobato, I. and Van Dyck, D. (2015). Multem: A new multislice program to perform accurate and fast electron diffraction and imaging simulations using graphics processing units with cuda. *Ultramicroscopy*, 156:9–17.

Longley, L., Collins, S. M., Zhou, C., Smales, G. J., Norman, S. E., Brownbill, N. J., Ashling, C. W., Chater, P. A., Tovey, R., Schönlieb, C.-B., et al. (2018). Liquid phase blending of metal-organic frameworks. *Nature Communications*, 9(1):1–10.

Lustig, M., Donoho, D., and Pauly, J. M. (2007). Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195.

Lyra, M. and Ploussi, A. (2011). Filtering in spect image reconstruction. *International Journal of Biomedical Imaging*, 2011.

Mahr, C., Müller-Caspary, K., Grieb, T., Schowalter, M., Mehrtens, T., Krause, F. F., Zillmann, D., and Rosenauer, A. (2015). Theoretical study of precision and accuracy of strain analysis by nano-beam electron diffraction. *Ultramicroscopy*, 158:38–48.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing.* Elsevier.

Merlin, T., Stute, S., Benoit, D., Bert, J., Carlier, T., Comtat, C., Filipovic, M., Lamare, F., and Visvikis, D. (2018). Castor: a generic data organization and processing code framework for multi-modal and multi-dimensional tomographic reconstruction. *Physics in Medicine & Biology*, 63(18):185005.

Midgley, P. A. and Weyland, M. (2003). 3d electron microscopy in the physical sciences: the development of z-contrast and eftem tomography. *Ultramicroscopy*, 96(3-4):413–431.

Monard, F., Nickl, R., Paternain, G. P., et al. (2019). Efficient nonparametric bayesian inference for $x$-ray transforms. *The Annals of Statistics*, 47(2):1113–1147.

Moriya, T., Saur, M., Stabrin, M., Merino, F., Voicu, H., Huang, Z., Penczek, P. A., Raunser, S., and Gatsogiannis, C. (2017). High-resolution single particle analysis from electron cryo-microscopy images using sphire. *Journal of Visualized Experiments.*

Mosek ApS (2010). The Mosek optimization software. Online at http://www.mosek.com.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., and Vagin, A. A. (2011). Refmac5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, 67(4):355–367.

Natterer, F. (2001). *The Mathematics of Computerized Tomography.* SIAM.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course.* Kluwer Academic Publishers Boston, Dordrecht, London.

Newton, M. C., Leake, S. J., Harder, R., and Robinson, I. K. (2010). Three-dimensional imaging of strain in a single zno nanorod. *Nature Materials*, 9(2):120–124.

Novikov, R. and Sharafutdinov, V. (2007). On the problem of polarization tomography: I. *Inverse Problems*, 23(3):1229–1257.

Novikov, R. G. (2002). An inversion formula for the attenuated x-ray transformation. *Arkiv för matematik*, 40(1):145–167.

Ochs, P., Chen, Y., Brox, T., and Pock, T. (2014). iPiano: Inertial Proximal Algorithm for Non-Convex Optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419.

Ochs, P., Fadili, J., and Brox, T. (2019). Non-smooth non-convex bregman minimization: Unification and new algorithms. *Journal of Optimization Theory and Applications*, 181(1):244–278.

Ophus, C. (2019). Four-Dimensional Scanning Transmission Electron Microscopy (4D-STEM): From Scanning Nanodiffraction to Ptychography and Beyond. *Microscopy and Microanalysis*, 25(3):563–582.

Ovesnỳ, M., Křížek, P., Borkovec, J., Švindrych, Z., and Hagen, G. M. (2014). Thunderstorm: a comprehensive imagej plug-in for palm and storm data analysis and super-resolution imaging. *Bioinformatics*, 30(16):2389–2390.

Paternain, G. P., Salo, M., and Uhlmann, G. (2014). Tensor tomography: progress and challenges. *Chinese Annals of Mathematics, Series B*, 35(3):399–428.

Pekin, T. C., Gammer, C., Ciston, J., Minor, A. M., and Ophus, C. (2017). Optimizing disk registration algorithms for nanobeam electron diffraction strain mapping. *Ultramicroscopy*, 176:170–176.

Pekin, T. C., Gammer, C., Ciston, J., Ophus, C., and Minor, A. M. (2018). In situ nanobeam electron diffraction strain mapping of planar slip in stainless steel. *Scripta Materialia*, 146:87–90.

Pelt, D. M. and Batenburg, K. J. (2013). Fast tomographic reconstruction from limited data using artificial neural networks. *IEEE Transactions on Image Processing*, 22(12):5238–5251.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). Ucsf chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612.

Pfeifer, M. A., Williams, G. J., Vartanyants, I. A., Harder, R., and Robinson, I. K. (2006). Three-dimensional mapping of a deformation field inside a nanocrystal. *Nature*, 442(7098):63–66.

Pock, T. and Sabach, S. (2016). Inertial Proximal Alternating Linearized Minimization (iPALM) for Nonconvex and Nonsmooth Problems. *SIAM Journal on Imaging Sciences*, 9(4):1756–1787.

Poon, C., Keriven, N., and Peyré, G. (2018). The geometry of off-the-grid compressed sensing. *arXiv preprint arXiv:1802.08464*.

Quinto, E. T. (1993). Singularities of the x-ray transform and limited data tomography in rˆ2 and rˆ3. *SIAM Journal on Mathematical Analysis*, 24(5):1215–1225.

Radon, J. (1917). Über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten. *Mathematische–Physikalische Klasse*, 69(4):262–277.

Righetto, R. D., Biyani, N., Kowal, J., Chami, M., and Stahlberg, H. (2019). Retrieving high-resolution information from disordered 2d crystals by single-particle cryo-em. *Nature Communications*, 10(1):1–10.

Robinson, I. and Harder, R. (2009). Coherent x-ray diffraction imaging of strain at the nanoscale. *Nature Materials*, 8(4):291–298.

Rouviere, J.-L., Béché, A., Martin, Y., Denneulin, T., and Cooper, D. (2013). Improved strain precision with high spatial resolution using nanobeam precession electron diffraction. *Applied Physics Letters*, 103(24):241913.

Rudin, L., Lions, P.-L., and Osher, S. (2003). Multiplicative denoising and deblurring: Theory and algorithms. In *Geometric level set methods in imaging, vision, and graphics*, pages 103–119. Springer.

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268.

Sage, D., Kirshner, H., Pengo, T., Stuurman, N., Min, J., Manley, S., and Unser, M. (2015). Quantitative evaluation of software packages for single-molecule localization microscopy. *Nature Methods*, 12(8):717–724.

Schermelleh, L., Ferrand, A., Huser, T., Eggeling, C., Sauer, M., Biehlmaier, O., and Drummen, G. P. (2019). Super-resolution microscopy demystified. *Nature cell biology*, 21(1):72–84.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682.

Schlom, D. G., Chen, L.-Q., Eom, C.-B., Rabe, K. M., Streiffer, S. K., and Triscone, J.-M. (2007). Strain Tuning of Ferroelectric Thin Films. *Annual Review of Materials Research*, 37(1):589–626.

Schmidt, M., Roux, N. L., and Bach, F. R. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, pages 1458–1466.

Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. part b–on the problem of osculatory interpolation. a second class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4(2):112–141.

Schwartz, M. and Shaw, L. (1975). *Signal Processing: Discrete Spectral Analysis, Detection, and Estimation.* New York: McGraw-Hill.

Sharafutdinov, V. A. (1994). *Integral Geometry of Tensor Fields.* De Gruyter, Berlin.

Smith, D. R. (1974). *Variational Methods in Optimization.* Prentice-Hall.

Solmon, D. C. (1976). The x-ray transform. *Journal of Mathematical Analysis and Applications*, 56(1):61 – 83.

Spall, J. C. (2005). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, volume 65. John Wiley & Sons.

Spitzbarth, M. and Drescher, M. (2015). Simultaneous iterative reconstruction technique software for spectral-spatial EPR imaging. *Journal of Magnetic Resonance*, 257:79–88.

Starborg, T., O'Sullivan, J. D., Carneiro, C. M., Behnsen, J., Else, K. J., Grencis, R. K., and Withers, P. J. (2019). Experimental steering of electron microscopy studies using prior x-ray computed tomography. *Ultramicroscopy*, 201:58–67.

Strang, G. (1972). Approximation in the finite element method. *Numerische Mathematik*, 19(1):81–98.

Stuart, A. M. (2010). Inverse problems: a bayesian perspective. *Acta Numerica*, 19:451.

Sullivan, T. J. (2015). *Bayesian Inverse Problems*, pages 91–112. Springer International Publishing, Cham.

Tao, S., Boley, D., and Zhang, S. (2016). Local linear convergence of ista and fista on the lasso problem. *SIAM Journal on Optimization*, 26(1):313–336.

Thirion, J.-P. (1991). A Geometric Alternative to Computed Tomography. In *Engineering in Medicine and Biology Society*, volume 13, page 34.

Tovey, R., Benning, M., Brune, C., Lagerwerf, M. J., Collins, S. M., Leary, R. K., Midgley, P. A., and Schönlieb, C.-B. (2019). Directional sinogram inpainting for limited angle tomography. *Inverse Problems*, 35(2):024004.

Tovey, R., Johnstone, D. N., Collins, S. M., Lionheart, W. R., Midgley, P. A., Benning, M., and Schönlieb, C.-B. (2020). Scanning electron diffraction tomography of strain. *arXiv preprint arXiv:2008.03281*.

Tovey, R. and Liang, J. (2020). The fun is finite: Douglas-rachford and sudoku puzzle–finite termination and local linear convergence. *arXiv preprint arXiv:2009.04018*.

Unser, M. (2019). A representer theorem for deep neural networks. *Journal of Machine Learning Research*, 20(110):1–30.

Unser, M. and Blu, T. (2003). Mathematical properties of the jpeg2000 wavelet filters. *IEEE Transactions on Image Processing*, 12(9):1080–1090.

Unser, M., Fageot, J., and Gupta, H. (2016). Representer theorems for sparsity-promoting $\ell^1$ regularization. *IEEE Transactions on Information Theory*, 62(9):5167–5180.

Usuda, K., Numata, T., Irisawa, T., Hirashita, N., and Takagi, S. (2005). Strain characterization in SOI and strained-Si on SGOI MOSFET channel using nano-beam electron diffraction (NBD). *Materials Science and Engineering: B*, 124-125:143–147.

Van Aarle, W., Palenstijn, W. J., Cant, J., Janssens, E., Bleichrodt, F., Dabravolski, A., De Beenhouwer, J., Batenburg, K. J., and Sijbers, J. (2016). Fast and flexible x-ray tomography using the astra toolbox. *Optics Express*, 24(22):25129–25147.

Vilas, J. L., Tagare, H. D., Vargas, J., Carazo, J. M., and Sorzano, C. O. S. (2020). Measuring local-directional resolution and local anisotropy in cryo-em maps. *Nature Communications*, 11(1):1–7.

Villa, S., Salzo, S., Baldassarre, L., and Verri, A. (2013). Accelerated and inexact forward-backward algorithms. *SIAM Journal on Optimization*, 23(3):1607–1633.

Vincent, R. and Midgley, P. (1994). Double conical beam-rocking system for measurement of integrated electron diffraction intensities. *Ultramicroscopy*, 53(3):271–282.

Weickert, J. (1998). *Anisotropic Diffusion in Image Processing*, volume 1. Teubner Stuttgart.

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34.

Xu, R., Chen, C.-C., Wu, L., Scott, M., Theis, W., Ophus, C., Bartels, M., Yang, Y., Ramezani-Dakhel, H., Sawaya, M. R., et al. (2015). Three-dimensional coordinates of individual atoms in materials revealed by electron tomography. *Nature Materials*, 14(11):1099–1103.

Zeltmann, S. E., Müller, A., Bustillo, K. C., Savitzky, B., Hughes, L., Minor, A. M., and Ophus, C. (2020). Patterned probes for high precision 4D-STEM bragg measurements. *Ultramicroscopy*, 209:112890.

Zhang, H., Wang, L., Duan, Y., Li, L., Hu, G., and Yan, B. (2017). Euler's Elastica Strategy for Limited-Angle Computed Tomography Image Reconstruction. *IEEE Transactions on Nuclear Science*.

Zhang, Y., Pu, Y. F., Hu, J. R., Liu, Y., and Zhou, J. L. (2011). A new CT metal artifacts reduction algorithm based on fractional-order sinogram inpainting. *Journal of X-Ray Science and Technology*, 19(3):373–384.

Zhao, G., Perilla, J. R., Yufenyuy, E. L., Meng, X., Chen, B., Ning, J., Ahn, J., Gronenborn, A. M., Schulten, K., Aiken, C., and Zhang, P. (2013). Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, 497(7451):643–646.

Ziemer, W. P. (1989). *Weakly Differentiable Functions. Sobolev Spaces and Function of Bounded Variation*, volume 120. Springer-Verlag New York.

Zivanov, J., Nakane, T., and Scheres, S. H. (2019). A bayesian approach to beam-induced motion correction in cryo-em single-particle analysis. *IUCrJ*, 6(1):5–17.

# Appendix A

# Supplementary Content for Limited Angle Tomography

## A.1 Theorem 2.2.1

**Theorem.** *If*

1. *$c_i$ are $2k$ times continuously differentiable in $\Delta$ and $\Sigma$, $k \geq 1$,*
2. *$c_1(\boldsymbol{x}|0,\Sigma) = c_2(\boldsymbol{x}|0,\Sigma)$ for all $x$ and $\Sigma \geq 0$,*
3. *and $\nabla_\Delta^{2j-1} c_1(\boldsymbol{x}|0,\Sigma) = \nabla_\Delta^{2j-1} c_2(\boldsymbol{x}|0,\Sigma) = 0$ for all $\boldsymbol{x}$ and $\Sigma \geq 0, j = 1\ldots,k$,*

*then $\mathcal{B}_\nu$ is $C^{2k-1}$ with respect to $\nu$ for all $\rho > 0, \sigma \geq 0$.*

In this proof we will drop the $\boldsymbol{x}$ argument from $c_i$ for conciseness of notation. Define

$$M_\nu = (\nabla \nu_\rho \nabla \nu_\rho^\top)_\sigma.$$

Note that convolutions are bounded linear maps and $\nabla\nu_\rho \in L^2$ by Young's inequality hence $M \colon L^1(\mathbb{R}^2,\mathbb{R}) \to L^\infty(\mathbb{R}^2, \mathrm{Sym}_2)$ is well defined and differentiable w.r.t. $\nu$. Hence, it suffices to show that $\mathcal{B}$ is differentiable w.r.t. $M$ where

$$M_\nu = \lambda_1 \boldsymbol{e}_1\boldsymbol{e}_1^\top + \lambda_2\boldsymbol{e}_2\boldsymbol{e}_2^\top, \lambda_1 \geq \lambda_2 \implies \mathcal{B} = c_1(\Delta,\Sigma)\boldsymbol{e}_1\boldsymbol{e}_1^\top + c_2(\Delta,\Sigma)\boldsymbol{e}_2\boldsymbol{e}_2^\top$$

and $\Delta = \lambda_1 - \lambda_2, \Sigma = \lambda_1 + \lambda_2$. Note that this is not a trivial statement, we can envisage very simple cases in which the (ordered) eigenvalue decomposition is not even continuous. For instance

$$M(t) = \begin{pmatrix} 1-t & 0 \\ 0 & t \end{pmatrix} \implies \Sigma(t) = 1, \Delta(t) = |1-2t|, \boldsymbol{e}_1 = \left\{ \begin{array}{ll} (1,0)^\top & t < 1/2 \\ (0,1)^\top & t > 1/2 \end{array} \right. .$$

This shows that, despite having $M \in C^\infty$, we cannot even guarantee that the decomposition is continuous and so cannot employ any chain rule to say that $\mathcal{B}$ is smooth.
The structure of this proof breaks into 4 parts:

1. If $c_1(0,\Sigma) = c_2(0,\Sigma)$, then $\mathcal{B}$ is well defined
2. If $c_i \in C^2$ for some open neighbourhood of point $\boldsymbol{x}$ such that $\lambda_1(\boldsymbol{x}) > \lambda_2(\boldsymbol{x})$, then $\mathcal{B}$ is differentiable w.r.t. $M$ on an open neighbourhood of $\boldsymbol{x}$
3. If $\nabla_\Delta c_1(0,\Sigma) = \nabla_\Delta c_2(0,\Sigma) = 0$ at a point, $\boldsymbol{x}$, where $\lambda_1(\boldsymbol{x}) = \lambda_2(\boldsymbol{x})$, then $\mathcal{B}$ is differentiable on an open neighbourhood of $\boldsymbol{x}$
4. If $\nabla_\Delta^{2j-1} c_1(0,\Sigma) = \nabla_\Delta^{2j-1} c_2(0,\Sigma) = 0$ at a point $\boldsymbol{x}$ where $\lambda_1(\boldsymbol{x}) = \lambda_2(\boldsymbol{x})$ and for all $j = 1\ldots,k$, then $\mathcal{B}$ is $C^{2k-1}$ on an open neighbourhood of $\boldsymbol{x}$

*Proof part 1.* Note that when $\lambda_1 = \lambda_2$ the choice of $\boldsymbol{e}_i$ is non-unique subject to $\boldsymbol{e}_1\boldsymbol{e}_1^\top + \boldsymbol{e}_2\boldsymbol{e}_2^\top = \mathrm{id}$ and so we get

$$\mathcal{B} = c_1(0,\Sigma)\,\mathrm{id} + (c_2(0,\Sigma) - c_1(0,\Sigma))\boldsymbol{e}_2\boldsymbol{e}_2^\top.$$

Therefore $\mathcal{B}$ is well defined if and only if $c_1(0,\Sigma) = c_2(0,\Sigma)$ for all $\Sigma \geq 0$.

As we are decomposing $2 \times 2$ matrices, it will simplify the proof to write explicit forms for the values under consideration:

$$M = \begin{pmatrix} m_{11} & m_{12} \\ m_{12} & m_{22} \end{pmatrix} \implies \lambda_i = \frac{m_{11} + m_{22} \pm \sqrt{(m_{11} - m_{22})^2 + 4m_{12}^2}}{2},$$

$$\Sigma = m_{11} + m_{22}, \quad \Delta = \sqrt{(m_{11} - m_{22})^2 + 4m_{12}^2},$$

$$\Delta \neq 0 \implies \boldsymbol{e}_1 = \frac{(2m_{12}, \Delta - m_{11} + m_{22})^\top}{\sqrt{(\Delta - m_{11} + m_{22})^2 + 4m_{12}^2}} = \frac{(\Delta + m_{11} - m_{22}, 2m_{12})^\top}{\sqrt{(\Delta + m_{11} - m_{22})^2 + 4m_{12}^2}},$$

$$\boldsymbol{e}_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \boldsymbol{e}_1.$$

We give two equations for $\boldsymbol{e}_1$ to account for the case when we get $\frac{(0,0)^\top}{0}$. □

*Proof part 2.* Note that $\Sigma$ is always smooth and $\Delta$ is smooth on the set $\{\Delta > 0\}$.

*Case $m_{12}(\boldsymbol{x}) \neq 0$:* Now both equations of $\boldsymbol{e}_1$ are valid (and equal) and the denominators non-zero on a neighbourhood. Hence, we can apply the standard chain rule and product rule to conclude.

*Case $m_{12}(\boldsymbol{x}) = 0$:* In this case $M(\boldsymbol{x})$ is diagonal but as $\Delta = |m_{11} - m_{22}| > 0$, we know that one of our formulae for $\boldsymbol{e}_1$ must be valid with non-vanishing denominator in a neighbourhood and so we can conclude as in the first case.

□

*Proof part 3.* Given the Neumann condition on $c_i$, we shall express $c_i$ locally by Taylor's theorem,

$$c_i(\Delta,\Sigma) = c_i(0,\Sigma) + O(\Delta^2) = c_1(0,\Sigma) + O(\Delta^2).$$

Now we consider a perturbation:

$$M = \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{12} & \varepsilon_{22} \end{pmatrix}$$

$$\begin{aligned} \implies \mathcal{B}(M+\varepsilon) - \mathcal{B}(M) &= (c_1(0, 2m+\varepsilon_{11}+\varepsilon_{22}) - c_1(0,2m))\operatorname{id} + O(\Delta^2), \\ \Delta^2 = (\varepsilon_{11}-\varepsilon_{22})^2 + 4\varepsilon_{12}^2 &= O(\|\varepsilon\|^2) \implies O(\Delta^2) \leq O(\|\varepsilon\|^2), \\ \implies \frac{\mathcal{B}(M+\varepsilon)-\mathcal{B}(M)}{\|\varepsilon\|} &= \frac{\nabla_\Sigma c_1(0,2m)\operatorname{tr}(\varepsilon)}{\|\varepsilon\|} + O(\|\varepsilon\|). \end{aligned}$$

In particular, $\mathcal{B}$ is differentiable w.r.t. $M$ at $\boldsymbol{x}$. □

*Proof part 4.* The proof of this follows exactly as the previous part,

$$c_i(\Delta,\Sigma) = \sum_0^{k-1} \frac{\Delta^{2j}}{j!} \nabla_\Delta^{2j} c_i(0,\Sigma) + O(\Delta^{2k})$$

where the remainder term is sufficiently smooth. Hence $c_i$ is at least $2k-1$ times differentiable with respect to $M$. □

## A.2 Theorem 2.3.1

**Theorem.** *If*

- *$c_i$ are bounded away from 0,*
- *$\rho > 0$,*
- *and $\mathcal{B}_d$ is differentiable in $d$,*

*then sublevel sets of* $\mathrm{E}$ *are weakly compact in* $L^2(\Omega,\mathbb{R}) \times L^2(\mathbb{R}^2,\mathbb{R})$ *and* $\mathrm{E}$ *is weakly lower semi-continuous. i.e. for all* $(u_n,v_n) \in L^2(\Omega,\mathbb{R}) \times L^2(\mathbb{R}^2,\mathbb{R})$*:*

$$\mathrm{E}(u_n,v_n) \textit{ uniformly bounded implies a subsequence converges weakly,}$$
$$\textit{and } \liminf_{n\to\infty} \mathrm{E}(u_n,v_n) \geq \mathrm{E}(u,v) \textit{ whenever } u_n \rightharpoonup u, v_n \rightharpoonup v.$$

*Proof.* If $c_i$ are bounded away from 0, then in particular we have $\mathcal{B}_{\mathcal{R}u_n} \gtrsim 1$ so $\mathrm{DTV}_{u_n}(v_n) = \|A_{\mathcal{R}u_n}\nabla v_n\| \gtrsim \|\nabla v_n\| = \mathrm{TV}(v_n)$. If $\{\mathrm{E}(u_n, v_n)\}$ is uniformly then, by definition,

$$\left\{\|\mathcal{S}_{\Omega'^c}(\mathcal{R}u_n - v_n)\|_2^2 + \|\mathcal{S}_{\Omega'}\mathcal{R}u_n - \eta\|_2^2 + \|\mathcal{S}_{\Omega'}v_n - \eta\|_2^2 + \mathrm{TV}(u_n) + \mathrm{TV}(v_n) \text{ s.t. } n \in \mathbb{N}\right\}$$

is also bounded. We conclude that $\left\{\left\|\mathcal{B}(u,v)^\top - \eta\right\|_2^2 + \mathrm{TV}\left((u,v)\right)\right\}$ is also uniformly bounded for some linear $\mathcal{B}$ and constant $\eta$.

Because of this, we are in a very classical setting where weak compactness can be shown in both the $\|(u,v)\|_2$ norm and $\|(u,v)\|_1 + \mathrm{TV}((u,v))$ (Chambolle and Lions, 1997).

We now proceed to the second claim, that E is weakly lower semi-continuous. Note that all of the convex terms in our energy are lower semi-continuous by classical arguments so it remains to show that the non-convex term is too. i.e.

$$(u_n, v_n) \rightharpoonup (u, v) \overset{?}{\implies} \liminf_{n\to\infty} \|\mathcal{B}_{\mathcal{R}u_n}\nabla v_n\|_{2,1} \geq \|\mathcal{B}_{\mathcal{R}u}\nabla v\|_{2,1}.$$

We shall first present a lemma from distribution theory.

**Lemma A.2.1.** *If $\Omega$ is compact, $\varphi \in C^\infty(\Omega, \mathbb{R})$ and $w_n \overset{L^p}{\rightharpoonup} w$, then*

$$w_n \star \varphi \to w \star \varphi \textit{ convergence in } C^k(\Omega, \mathbb{R}) \textit{ for all } k < \infty.$$

*Proof.* Recall that $w_n \rightharpoonup w \implies \|w_n\|_p \leq W$ for some $W < \infty$. By Hölder's inequality:

$$|w_n \star \varphi(\boldsymbol{x}) - w \star \varphi(\boldsymbol{y})| \leq \int |w_n(\boldsymbol{z})||\varphi(\boldsymbol{x} - \boldsymbol{z}) - \varphi(\boldsymbol{y} - \boldsymbol{z})| \lesssim_{p,\Omega} |\boldsymbol{x} - \boldsymbol{y}| W \|\varphi'\|_\infty,$$

therefore

$$|w_n \star \varphi(\boldsymbol{x})| \lesssim_{p,\Omega} W \|\varphi\|_\infty.$$

i.e. $\{w_n \text{ s.t. } n \in \mathbb{N}\}$ is uniformly bounded and uniformly (Lipschitz) continuous. We conclude

$$w_n \rightharpoonup w \implies w_n \star \varphi(\boldsymbol{x}) - w \star \varphi(\boldsymbol{x}) = \langle w_n - w, \ \varphi(\boldsymbol{x} - \cdot)\rangle \to 0 \text{ for every } \boldsymbol{x}.$$

Hence, we also know $w_n \star \varphi$ converges point-wise to a unique continuous function. Suppose $\|w_{n_k} \star \varphi - w \star \varphi\|_\infty \geq \varepsilon > 0$ for some $\varepsilon$ and subsequence $n_k \to \infty$. By the Arzela-Ascoli theorem, some further subsequence must converge uniformly to the point-wise limit, $w \star \varphi$, which gives us the required contradiction. Hence, $w_n \star \varphi \to w \star \varphi$ in $L^\infty = C^0$. The general theorem follows by induction. □

This lemma gives us two direct results:

$$\rho > 0 \implies (\mathcal{R}u_n)_\rho \to (\mathcal{R}u)_\rho \text{ in } L^\infty,$$

and

$$\{(\mathcal{R}u_n)_\rho\} \cup \{(\mathcal{R}u)_\rho\} \text{ compact, } \mathcal{B}_\nu \in C^1 \text{ w.r.t. } \nu \implies \mathcal{B}_{\mathcal{R}u_n} \to \mathcal{B}_{\mathcal{R}u} \text{ in } \|\cdot\|_{2,\infty}.$$

Hence, whenever $w \in W^{1,1}$ we have

$$\begin{aligned} \|\mathcal{B}_{\mathcal{R}u_n}\nabla w\| &\geq \|\mathcal{B}_{\mathcal{R}u}\nabla w\| - \|(\mathcal{B}_{\mathcal{R}u_n} - \mathcal{B}_{\mathcal{R}u})\nabla w\| \\ &\geq \|\mathcal{B}_{\mathcal{R}u}\nabla w\| - \|\mathcal{B}_{\mathcal{R}u_n} - \mathcal{B}_{\mathcal{R}u}\|_{2,\infty} \operatorname{TV}(w). \end{aligned}$$

By density of $W^{1,1}$ in the space of Bounded Variation, we know the same holds for $w = v_n$. We also know $\operatorname{TV}(v_n)$ is uniformly bounded, thus

$$\liminf_{n\to\infty} \|\mathcal{B}_{\mathcal{R}u_n}\nabla v_n\| = \liminf_{n\to\infty} \|\mathcal{B}_{\mathcal{R}u}\nabla v_n\|.$$

Hence, for all $\|\varphi\|_{2,\infty} \leq 1$ we have

$$\langle v,\ \nabla \cdot (\mathcal{B}_{\mathcal{R}u}\varphi)\rangle = \liminf_{n\to\infty} \langle v_n,\ \nabla \cdot (\mathcal{B}_{\mathcal{R}u}\varphi)\rangle \leq \liminf_{n\to\infty} \|\mathcal{B}_{\mathcal{R}u}\nabla v_n\| \leq \liminf_{n\to\infty} \|\mathcal{B}_{\mathcal{R}u_n}\nabla v_n\|$$

as required. □

## A.3 Sensitivity to hyperparameters

As has been noted in the main text, there are many hyper-parameters to tune for the best reconstruction. We commonly found that reconstructions were qualitatively insensitive near the optimal parameter choice, but we include here some illustrations of the typical effect of each parameter. To recap, the full model is

$$E(u,v) = \frac{1}{2}\|\mathcal{R}u - v\|_{\alpha_1}^2 + \frac{\alpha_2}{2}\|S\mathcal{R}u - \eta\|_2^2 + \frac{\alpha_3}{2}\|Sv - b\|_2^2 + \beta_1 \operatorname{TV}(u) + \beta_2 \operatorname{DTV}(v).$$

To remove a degree of freedom, we have always normalised such that $\alpha_2 = 1$. To construct the directional TV functional we need 2 smoothing parameters, $\rho$ and $\sigma$ defining

$$\begin{aligned} A_d(x) &\coloneqq c_1(\lambda_1(x),\lambda_2(x))\boldsymbol{e}_1(x)\boldsymbol{e}_1(x)^T + c_2(\lambda_1(x),\lambda_2(x))\boldsymbol{e}_2(x)\boldsymbol{e}_2(x)^T \\ \text{such that} \qquad &(\nabla d_\rho \nabla d_\rho^T)_\sigma(x) = \lambda_1(x)\boldsymbol{e}_1(x)\boldsymbol{e}_1(x)^T + \lambda_2(x)\boldsymbol{e}_2(x)\boldsymbol{e}_2(x)^T, \\ &\lambda_1(x) \geq \lambda_2(x) \geq 0. \end{aligned}$$

Again, we kept $\rho = 1$ fixed and only show reconstructions for different values of $\sigma$. The optimal parameters for the Shepp-Logan phantom referred to below were

$$\alpha_1 = \frac{1}{4^2}\mathbb{1}_{\Omega'^c},\ \alpha_3 = 3\times 10^{-1},\ \beta_1 = 3\times 10^{-5},\ \beta_2 = 3\times 10^{2},\ \beta_3 = 10^{10},\ \sigma = 8.$$

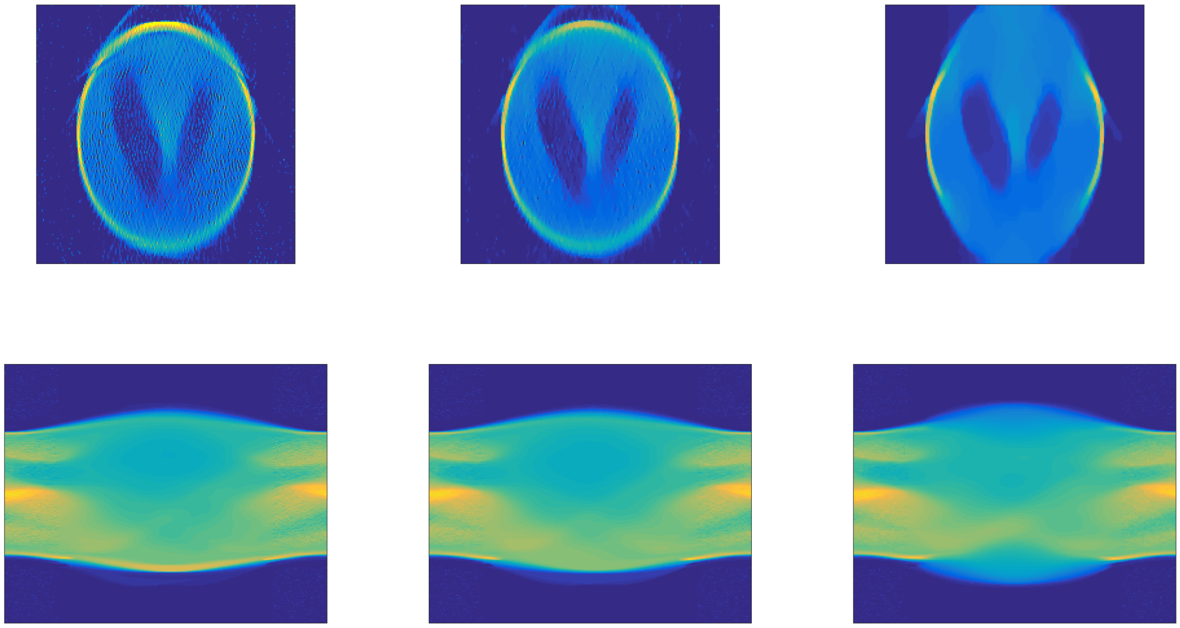


**Figure A.1** Varying reconstruction for low (first column), optimal (middle column) and high (right column) values of $\beta_1$ (TV regularisation parameter). ‘low’ is a factor of 0.1 lower than ‘optimal’ and ‘high’ a factor of 10 higher.

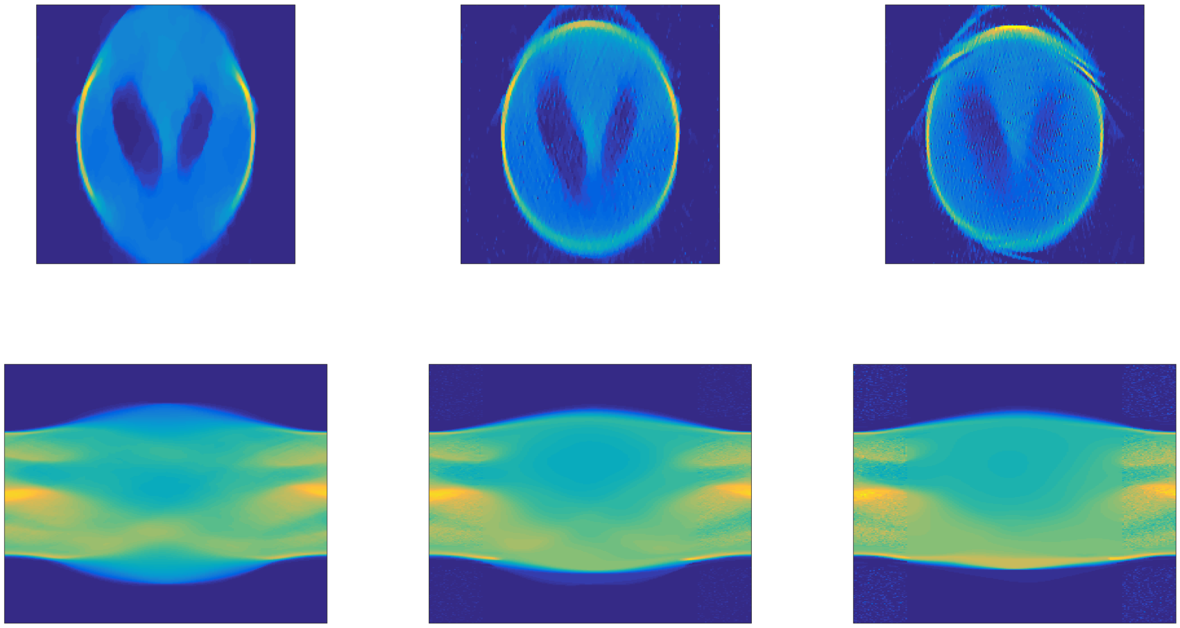


**Figure A.2** Varying reconstruction for low (first column), optimal (middle column) and high (right column) values of $\beta_2$ (DTV regularisation parameter). ‘low’ is a factor of 0.1 lower than ‘optimal’ and ‘high’ a factor of 10 higher.

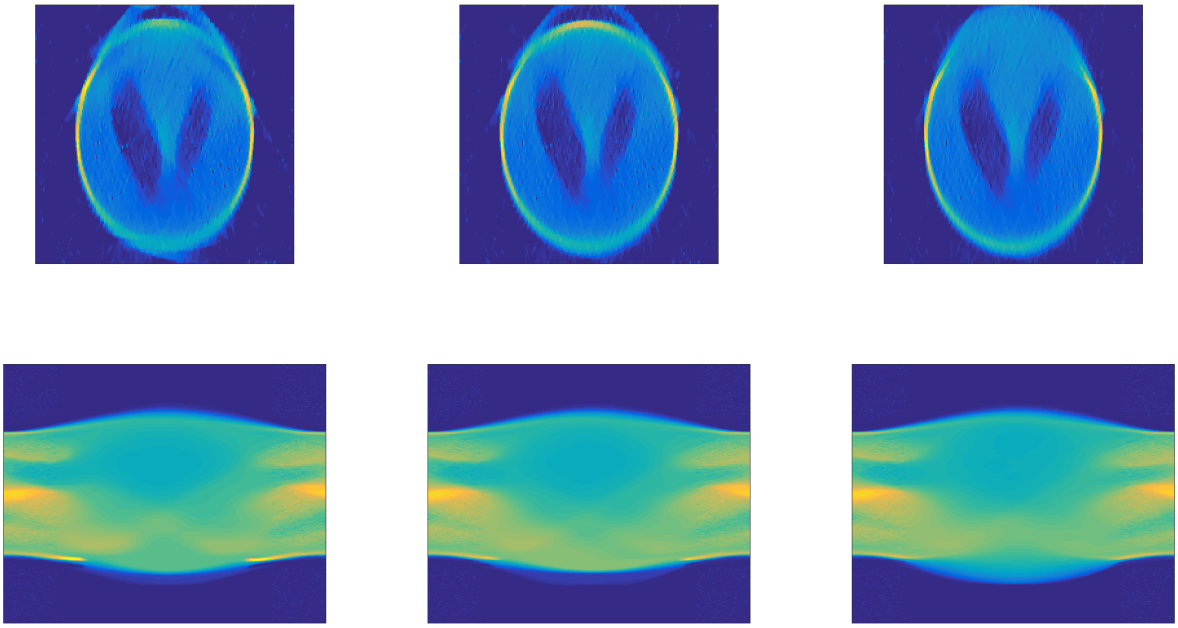


**Figure A.3** Varying reconstruction for low (first column), optimal (middle column) and high (right column) values of $\alpha_1$ (pairing term between $u$ and $v$). ‘low’ is a factor of 0.1 lower than ‘optimal’ and ‘high’ a factor of 10 higher.

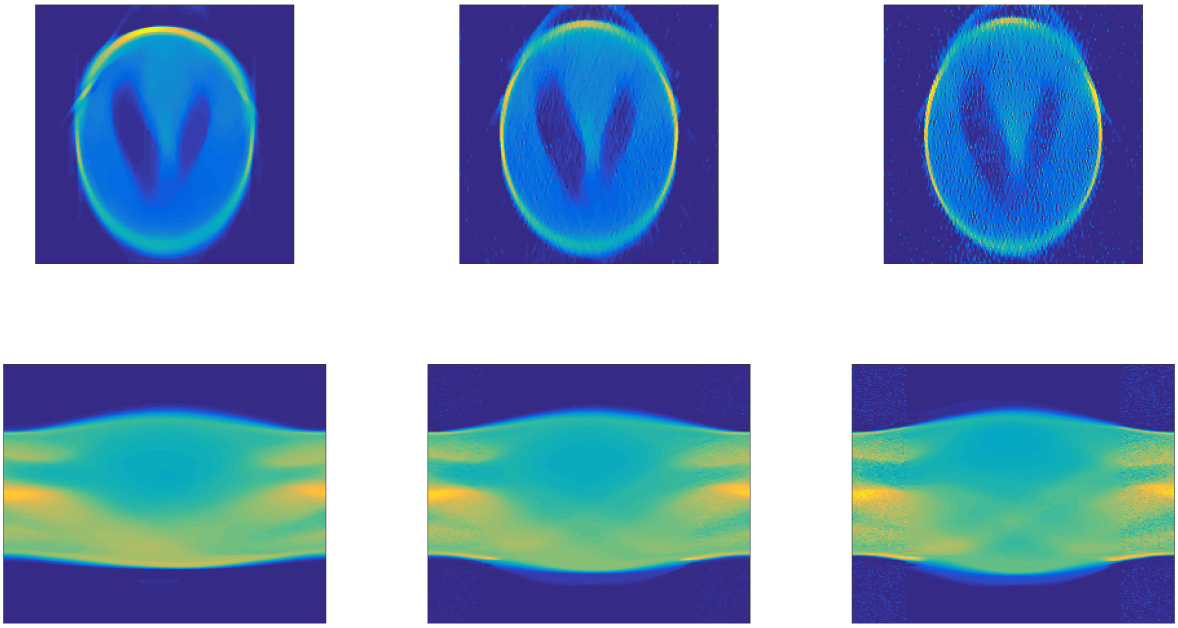


**Figure A.4** Varying reconstruction for low (first column), optimal (middle column) and high (right column) values of $\alpha_3$ (sinogram noise parameter). ‘low’ is a factor of 0.1 lower than ‘optimal’ and ‘high’ a factor of 10 higher.

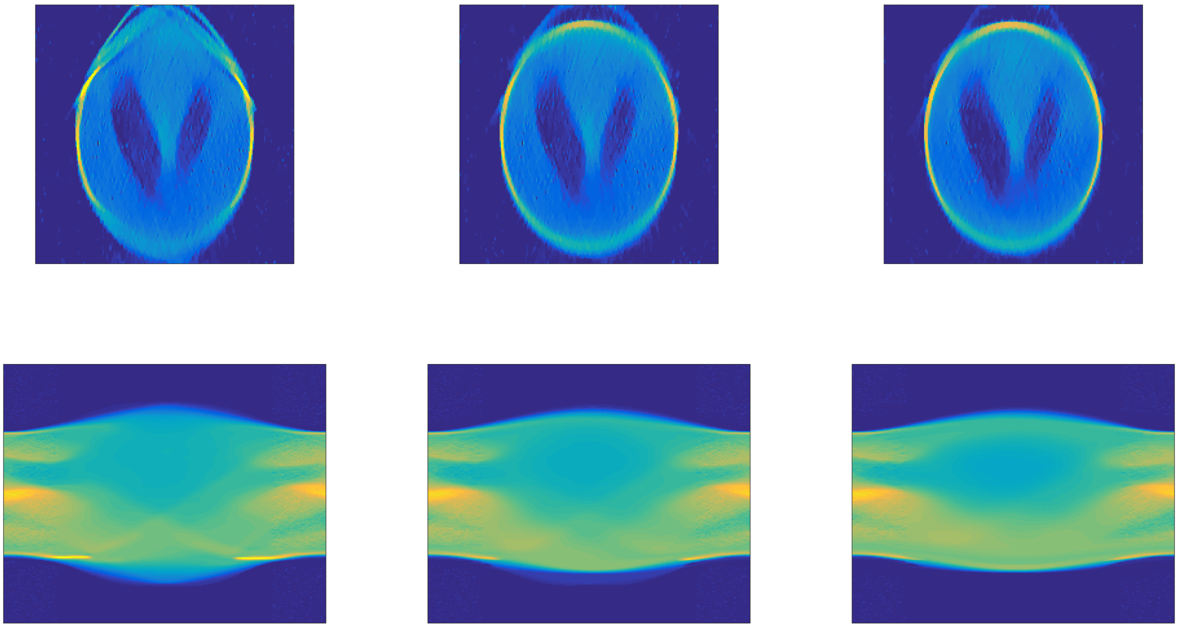


**Figure A.5** Varying reconstruction for low (first column), optimal (middle column) and high (right column) values of $\sigma$ (smoothing parameter inside DTV functional). 'low' is a factor of 0.5 lower than 'optimal' and 'high' a factor of 2 higher.

# Appendix B

# Supplementary Content for Strain Tomography of Crystals

## B.1 Theorem 3.3.2 and Lemma 3.3.3

**Theorem** (Theorem 3.3.2)**.** *If*

$$u'(\boldsymbol{r}) = u(\boldsymbol{r} + \vec{R}(\boldsymbol{r})) = u(A\boldsymbol{r} + \boldsymbol{b}) \text{ for some } A \in \mathbb{R}^{3\times 3}, \boldsymbol{b} \in \mathbb{R}^3, u \in L^2(\mathbb{R}^n; \mathbb{C}),$$

*where $A$ is invertible, then we can express its Fourier transform as*

$$\mathcal{F}[u'](\boldsymbol{K}) = \det(A)^{-1} e^{\imath \boldsymbol{b} \boldsymbol{\cdot} A^{-\top}\boldsymbol{K}} \mathcal{F}[u](A^{-\top}\boldsymbol{K}).$$

*Proof.* The proof is purely analytical:

$$\begin{aligned}
\mathcal{F}[u'](\boldsymbol{K}) &= \int_{\mathbb{R}^3} u'(\boldsymbol{r}) \exp\left(-\imath \boldsymbol{r} \boldsymbol{\cdot} \boldsymbol{K}\right) d\boldsymbol{r} \\
&= \int_{\mathbb{R}^3} u(A\boldsymbol{r} + \boldsymbol{b}) \exp\left(-\imath (A^{-1}A\boldsymbol{r}) \boldsymbol{\cdot} \boldsymbol{K}\right) d\boldsymbol{r} \\
&= \int_{\mathbb{R}^3} u(A\boldsymbol{r} + \boldsymbol{b}) \exp\left(-\imath (A\boldsymbol{r}) \boldsymbol{\cdot} (A^{-\top}\boldsymbol{K})\right) d\boldsymbol{r} \\
&= \int_{\mathbb{R}^3} u(\boldsymbol{r}' + \boldsymbol{b}) \exp\left(-\imath (\boldsymbol{r}' + \boldsymbol{b} - \boldsymbol{b}) \boldsymbol{\cdot} A^{-\top}\boldsymbol{K}\right) \frac{d\boldsymbol{r}'}{\det(A)} \\
&= \det(A)^{-1} \int_{\mathbb{R}^3} u(\boldsymbol{r}') \exp\left(-\imath \boldsymbol{r}' \boldsymbol{\cdot} A^{-\top}\boldsymbol{K} + \imath \boldsymbol{b} \boldsymbol{\cdot} A^{-\top}\boldsymbol{K}\right) d\boldsymbol{r}' \\
&= \det(A)^{-1} e^{\imath \boldsymbol{b} \boldsymbol{\cdot} A^{-\top}\boldsymbol{K}} \mathcal{F}[u](A^{-\top}\boldsymbol{K}).
\end{aligned}$$

□

**Theorem** (Lemma 3.3.3)**.** *If $A \in \mathbb{R}^{3\times 3}$ is an invertible matrix and $\boldsymbol{b} \in \mathbb{R}^3$, then*

$$\mathcal{F}[u_0(A\cdot +\boldsymbol{b})](\boldsymbol{K}) = e^{\imath \boldsymbol{b}\bullet A^{-\top}\boldsymbol{K}} \sum_{i=1}^{\infty} a_i \delta_{A^\top \boldsymbol{p}_i}(\boldsymbol{K}).$$

*Proof.* From Theorem 3.3.2, we know

$$\begin{aligned}\mathcal{F}[u_0(A\cdot +\boldsymbol{b})](\boldsymbol{K}) &= \det(A)^{-1} e^{\imath \boldsymbol{b}\bullet A^{-\top}\boldsymbol{K}} \mathcal{F}[u_0](A^{-\top}\boldsymbol{K}) \\ &= \sum_{i=1}^{\infty} a_i \det(A)^{-1} e^{\imath \boldsymbol{b}\bullet A^{-\top}\boldsymbol{K}} \delta_{\boldsymbol{p}_i}(A^{-\top}\boldsymbol{K}).\end{aligned}$$

Thus, to complete the lemma, it suffices to show

$$\det(A)^{-1}\delta_{\boldsymbol{p}}(A^{-\top}\boldsymbol{K}) = \delta_{A^\top \boldsymbol{p}}(\boldsymbol{K}).$$

This is verified by an arbitrary test function, $\varphi \in C_c^\infty$

$$\int_{\mathbb{R}^3} \det(A)^{-1}\delta_{\boldsymbol{p}}(A^{-\top}\boldsymbol{K})\varphi(\boldsymbol{K})d\boldsymbol{K} = \int_{\mathbb{R}^3} \delta_{\boldsymbol{p}}(\boldsymbol{K})\varphi(A^\top\boldsymbol{K})d\boldsymbol{K} = \varphi(A^\top\boldsymbol{p}).$$

□

## B.2 Probability background

We recap some basic concepts and technical results from probability theory which are needed in the proofs in Section 3.5.

**Definition B.2.1.**

- *$X\colon t \mapsto X_t \in \mathbb{C}$ can be called a* random variable *where random (complex) values of $X$ can be sampled by sampling indices $t \in [t_0, t_1]$ uniformly at random, i.e. $t \sim \text{Uniform}[t_0, t_1)$.*
- *For a random variable, $X$, we define its* expectation *to be*

$$\mathbb{E}\, X = \mathbb{E}_t\, X_t = \frac{\int_{t_0}^{t_1} X_t dt}{|t_1 - t_0|}.$$

*If $t$ is a discrete index, then integral can be replaced with summation and, for $t_0, t_1 \in \mathbb{Z}$, this becomes*

$$\mathbb{E}\, X = \mathbb{E}_t\, X_t = \frac{\sum_{t_0}^{t_1-1} X_t}{|t_1 - t_0|}.$$

*In quantum physics this would also be called the* expectation value*, $\mathbb{E}\, X = \langle X\rangle$.*

- *For a random variable $X$ and value $x \in \mathbb{C}$ we say denote the probability that $X_t = x$ to be*

$$\text{probability}(X = x) = \text{probability}(X_t = x) = \frac{|\{t \ s.t. \ X_t = x\}|}{|t_1 - t_0|}.$$

- *Let $X$ and $Y$ be random variables. For $x, y \in \mathbb{C}$, we say the events $X_t = x$ and $Y_t = y$ are* independent *if*

$$\text{probability}(X_t = x \ \textit{and} \ Y_t = y) = \text{probability}(X_t = x) \cdot \text{probability}(Y_t = y).$$

- *We say that two random variables, $X$ and $Y$, are* independent *if*

$$\textit{events } X_t = x \textit{ and } Y_t = y \textit{ are independent for all } x, y.$$

**Lemma B.2.2.** *A standard result from probability theory: If $X$ and $Y$ are two independent random variables, then*

$$\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y).$$

*Also, for any continuous function, $\varphi$, the random variable $\varphi(X)_{i,j,t} = \varphi(X_{i,j,t})$ is also independent of $Y$.*

**Lemma B.2.3.** *Let $\varphi \colon \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function and let $\vec{\eta} \colon [0,1] \to \mathbb{R}^n$ be a random variable, then*

$$\mathop{\mathbb{E}}_t \varphi(\vec{\eta}_t) = \varphi(\mathop{\mathbb{E}}_t \vec{\eta}) + O(\|\vec{\eta} - \mathbb{E}\,\vec{\eta}\|_\infty^2 \left\|\nabla^2\varphi\right\|_\infty)$$

*Proof.* Because the length of the path (size of domain of $\vec{\eta}$) is one, we can replace expectation with integration and then replace $\varphi$ with its standard Taylor expansion:

$$\begin{aligned}
\mathop{\mathbb{E}}_t \varphi(\vec{\eta}_t) &= \int_0^1 \varphi(\vec{\eta}_t)dt \qquad\qquad \rightsquigarrow \text{(Taylor expansion about } \mathbb{E}\,\vec{\eta})\\
&= \int_0^1 \varphi(\mathbb{E}\,\vec{\eta}) + (\vec{\eta}_t - \mathbb{E}\,\vec{\eta}) \bullet \nabla\varphi(\mathbb{E}\,\vec{\eta}) + O\left(|\vec{\eta}_t - \mathbb{E}\,\vec{\eta}|^2 \left\|\nabla^2\varphi\right\|_\infty\right) dt\\
&= \varphi(\mathbb{E}\,\vec{\eta}) + \left(\int_0^1 \vec{\eta}_t - \mathbb{E}\,\vec{\eta} dt\right) \bullet \nabla\varphi(\mathbb{E}\,\vec{\eta}) + O\left(\|\vec{\eta} - \mathbb{E}\,\vec{\eta}\|_\infty^2 \left\|\nabla^2\varphi\right\|_\infty\right)\\
&= \varphi(\mathbb{E}\,\vec{\eta}) + (\mathbb{E}[\vec{\eta}] - 1\,\mathbb{E}\,\vec{\eta}) \bullet \nabla\varphi(\mathbb{E}\,\vec{\eta}) + O\left(\|\vec{\eta} - \mathbb{E}\,\vec{\eta}\|_2^2 \left\|\nabla^2\varphi\right\|_\infty\right)
\end{aligned}$$

as required. □

## B.3 Precession angle estimation

Figure B.1 sketches the choice of precession angle for a deformed sample. There are two triangles of interest, one in the positive $z$ direction which accounts for the curvature of the sphere, and another in the negative direction to account for strain. The upper triangle is a right angled triangle with one corner at the origin and the other at a point where the sphere of radius $P$ meets the Ewald sphere, say at point $(k, k_z(k))$. This gives the relationship

$$\begin{aligned} P^2 &= k^2 + k_z(k)^2 \\ &= k^2 + 4\pi^2\lambda^{-2} + 4\pi^2\lambda^{-2} - k^2 - 4\pi\lambda^{-1}\sqrt{4\pi^2\lambda^{-2} - k^2} \\ &= 4\pi\lambda^{-1}k_z(k), \end{aligned}$$

i.e. $k_z(k) = \frac{\lambda P^2}{4\pi}$. On the other hand, the strain moves the point $(P, 0)$ a maximal distance of $\sigma P$ from its starting point and away from the Ewald sphere. Assuming the worst strain is a rotation, we get an isosceles triangle whose angle can be computed with the cosine rule:

$$\cos(\theta) = \frac{2P^2 - \sigma^2 P^2}{2P^2} = 1 - \frac{\sigma^2}{2}.$$

Combining, the maximal angle is

$$\alpha = \cos^{-1}\left(1 - \frac{\sigma^2}{2}\right) + \sin^{-1}\left(\frac{\lambda P}{4\pi}\right).$$

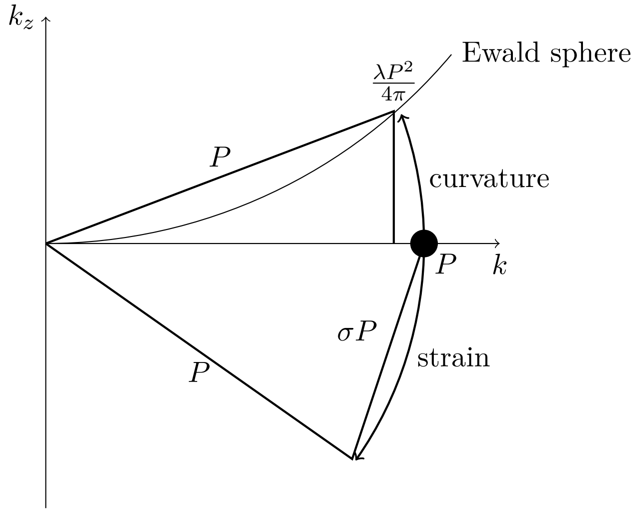


**Figure B.1** Geometrical argument for choosing precession angle

## B.4 Continuous deformation phantom

Examples of strained discs for the layered phantoms are given in Figure 3.3, however, the discs for the continuously deformed phantom looked qualitatively different. An example is given in Figure B.2.

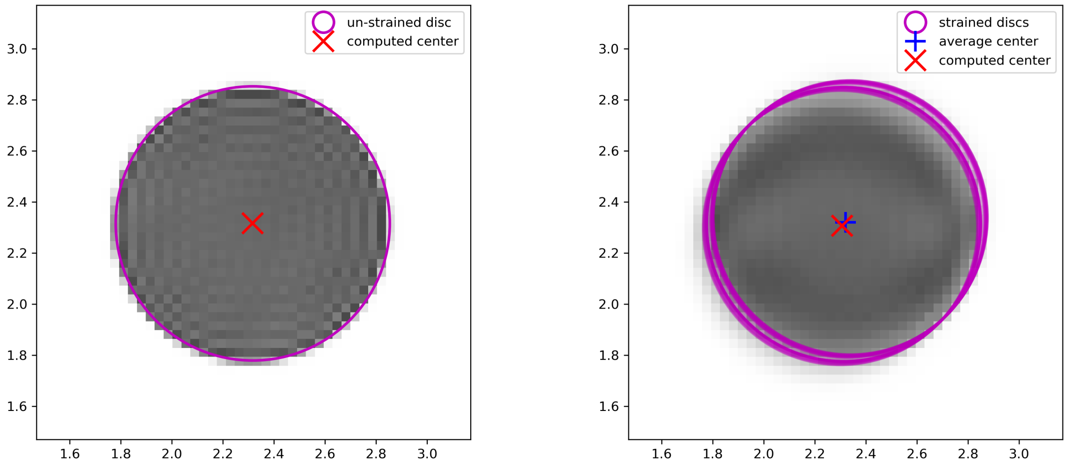


**Figure B.2** Example disc with worst centre of mass prediction error for the continuous deformation phantom. Left/right hand plots show un-strained/strained disc respectively.

## B.5 Precession discretisation

For the dynamical simulations, multiple simulations were run with different numbers of points discretising precession; Figure B.3 shows convergence to the corresponding values stated in Table 3.1. Dynamical results stated in Table 3.1 are with $2^8$ precession points. Because of the number of kinematical simulations, the full plots were not computed. Instead, as $\alpha = 2°$ was the slowest to converge for dynamical, the number of precession points was increased until the values stated in Table 3.1 for kinematical simulation at $\alpha = 2°$ had converged within the necessary two decimal place rounding error. Kinematical results stated in Table 3.1 are with $2^5$ precession points.

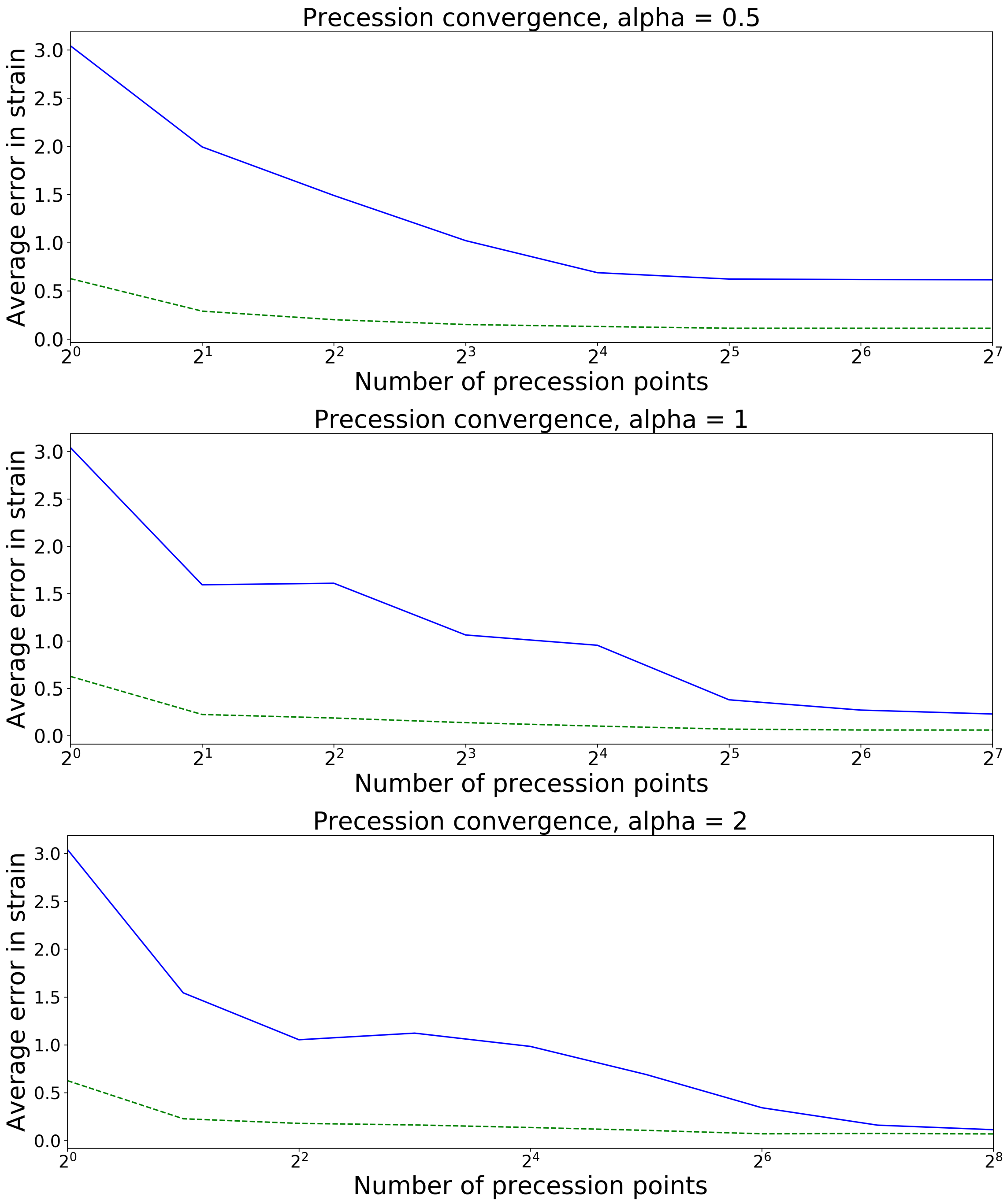


**Figure B.3** Convergence plot of the values in Table 3.1 for dynamical simulation for different numbers of precession points. The solid and dashed lines shows the convergence of centre of mass accuracy and registration method respectively.

## B.6 Reconstruction error plots

In Section 3.7.2 a reconstructed gradient deformation tensor is computed and the $99^{\text{th}}$ percentile errors are reported. Figure B.4 demonstrates the spread of errors in more detail. The first column represents the middle slice (a best-case analysis) which shows that errors can be up to a factor of four smaller than the maximum (third column). The $99^{\text{th}}$ percentile gives an average worst-case. Qualitatively, this agrees much more with the middle slice indicating that the largest errors are achieved on the interface voxels between crystal and vacuum and on the smooth interior errors are much lower, up to a factor of three.

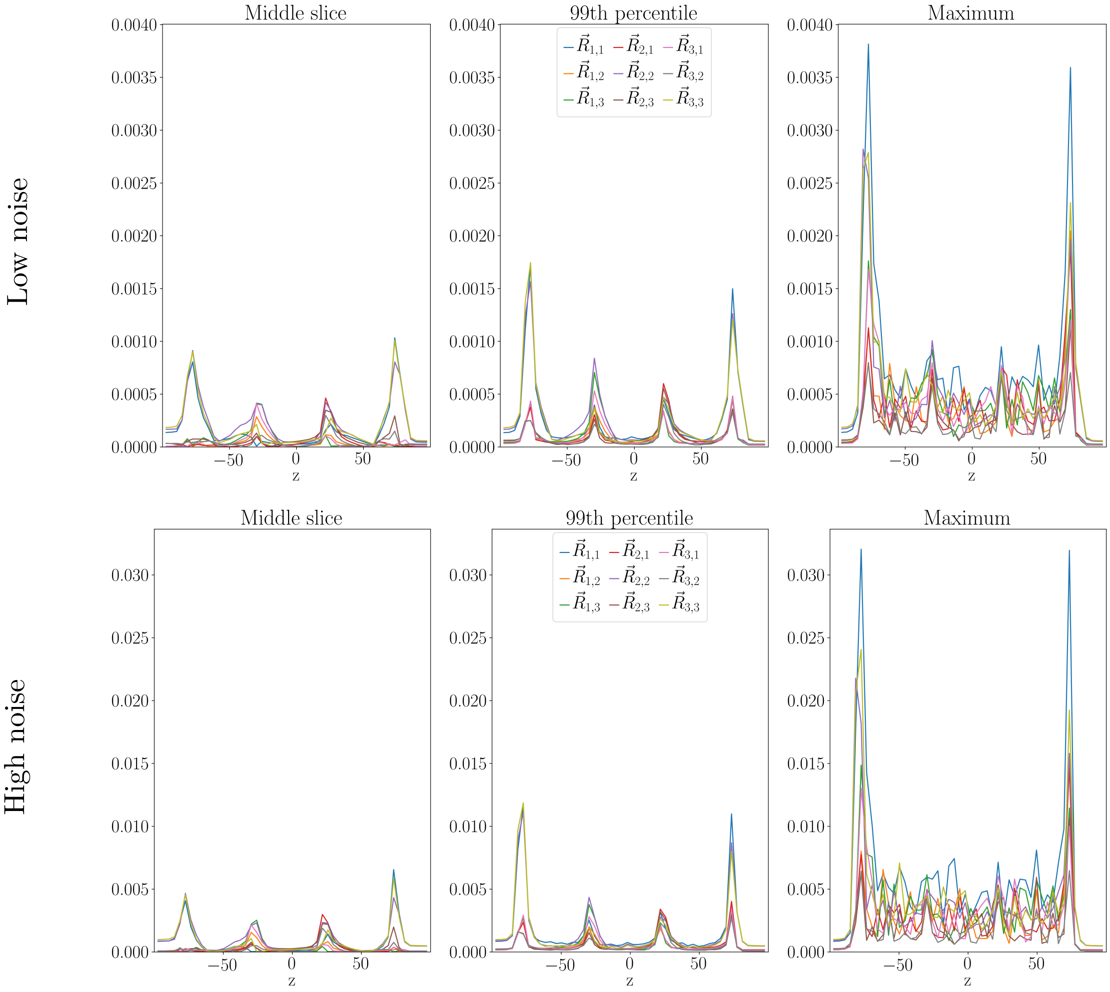


**Figure B.4** 1D projections of error of the gradient deformation tensor. For each 1D point shown, the corresponding 2D slice is projected to the reported error by the indicated method. The first pixel extracts the physically central pixel, the second computes the $99^{\text{th}}$ percentile and the final computes the maximum over all pixels.

# Appendix C

# Adaptive FISTA Appendices

## C.1 Proofs for FISTA convergence

This section contains all of the statements and proofs of the results contained in Section 4.4.

### C.1.1 Proofs for Step 2

**Lemma C.1.1** (Lemma 4.4.2)**.** *Suppose $\nabla \mathrm{f}$ is 1-Lipschitz, for any $\overline{u} \in \mathbb{U}^{n-1}$ define*

$$u := \operatorname*{argmin}_{u \in \mathbb{U}^n} \tfrac{1}{2}\left\|u - \overline{u} + \nabla \mathrm{f}(\overline{u})\right\|^2 + \mathrm{g}(u).$$

*Then, for all $w \in \overline{(\mathbb{U}^n)^*} \supset \mathbb{H}$, we have*

$$\mathrm{E}(u) + \tfrac{1}{2}\left\|u - \Pi_n w\right\|^2 \leq \mathrm{E}(\Pi_n w) + \tfrac{1}{2}\left\|\overline{u} - \Pi_n w\right\|^2$$

*where $\Pi_n \colon \overline{(\mathbb{U}^n)^*} \to \mathbb{U}^n$ is the orthogonal projection.*

*Proof.* This is exactly the result of (Chambolle and Dossal, 2015, Lemma 1) applied to the function $u \mapsto \mathrm{E}(\Pi_n u)$. □

---

**Theorem** (Lemma 4.4.3)**.** *Let $\mathbb{U}^n \subset \mathbb{H} \cap \mathbb{U}$ and $w_n \in \mathbb{U}^n$ be chosen arbitrarily and $u_n/v_n$ be generated by Algorithm 4.1 for all $n \in \mathbb{N}$. For all $n \in \mathbb{N}$ it holds that*

$$t_n^2(\mathrm{E}(u_n) - \mathrm{E}(w_n)) - (t_n^2 - t_n)(\mathrm{E}(u_{n-1}) - \mathrm{E}(w_n)) \leq \tfrac{1}{2}\left[\left\|v_{n-1}\right\|^2 - \left\|v_n\right\|^2\right] + \left\langle v_n - v_{n-1},\ w_n\right\rangle. \tag{C.1}$$

*Proof.* Modifying (Chambolle and Dossal, 2015, Theorem 2), for $n \geq 1$ we apply Lemma C.1.1 with $\overline{u} = \overline{u}_{n-1}$ and $w = (1 - \frac{1}{t_n})u_{n-1} + \frac{1}{t_n}w_n$. This gives

$$\mathrm{E}(u_n) + \tfrac{1}{2}\left\|\tfrac{1}{t_n}v_n - \tfrac{1}{t_n}w_n\right\|^2 \leq \mathrm{E}\left((1 - \tfrac{1}{t_n})u_{n-1} + \tfrac{1}{t_n}w_n\right) + \tfrac{1}{2}\left\|\tfrac{1}{t_n}v_{n-1} - \tfrac{1}{t_n}w_n\right\|^2.$$

By the convexity of E, this reduces to

$$\begin{aligned}\mathrm{E}(u_n)-\mathrm{E}(w_n)-(1-\tfrac{1}{t_n})[\mathrm{E}(u_{n-1})-\mathrm{E}(w_n)] &\leq \tfrac{1}{2t_n^2}\left\|v_{n-1}-w_n\right\|^2-\tfrac{1}{2t_n^2}\left\|v_n-w_n\right\|^2\\ &= \tfrac{1}{2t_n^2}\left[\left\|v_{n-1}\right\|^2-\left\|v_n\right\|^2\right]+\tfrac{1}{t_n^2}\left\langle v_n-v_{n-1},\ w_n\right\rangle.\end{aligned}$$

□

**Theorem C.1.2** (Theorem 4.4.4)**.** *Fix a sequence of subspaces* $\{\mathbb{U}^n\subset\mathbb{U}\cap\mathbb{H}\text{ s.t. }n\in\mathbb{N}\}$*, arbitrary* $u_0\in\mathbb{U}^0$*, and FISTA stepsize choice* $(t_n)_{n\in\mathbb{N}}$*. Let* $u_n$ *and* $v_n$ *be generated by Algorithm 4.1. Then, for any choice of* $w_n\in\mathbb{U}^n$ *and* $N\in\mathbb{N}$ *we have*

$$\begin{aligned}t_N^2\,\mathrm{E}_0(u_N)+\sum_{n=1}^{N-1}\rho_n\,\mathrm{E}_0(u_n)+\frac{\left\|v_N-w_N\right\|^2}{2}&\leq\frac{\left\|u_0-w_0\right\|^2-\left\|w_0\right\|^2+\left\|w_N\right\|^2}{2}\\ &\quad+\sum_{n=1}^{N}t_n\,\mathrm{E}_0(w_n)+\left\langle v_{n-1},\ w_{n-1}-w_n\right\rangle. \qquad (\mathrm{C.2})\end{aligned}$$

*Proof.* Theorem C.1.2 is just a summation of (C.1) over all $n=1,\dots,N$. To see this: first add and subtract $\mathrm{E}(u^*)$ to each term on the left-hand side to convert E to $\mathrm{E}_0$, then move $\mathrm{E}_0(w_n)$ to the right-hand side. Now (C.1) becomes

$$t_n^2\,\mathrm{E}_0(u_n)-(t_n^2-t_n)\,\mathrm{E}_0(u_{n-1})\leq t_n\,\mathrm{E}_0(w_n)+\tfrac{1}{2}\left[\left\|v_{n-1}\right\|^2-\left\|v_n\right\|^2\right]+\left\langle v_n-v_{n-1},\ w_n\right\rangle.$$

Summing this inequality from $n=1$ to $n=N$ gives

$$t_N^2\,\mathrm{E}_0(u_N)+\sum_{n=1}^{N-1}\underbrace{(t_n^2-t_{n+1}^2+t_{n+1})}_{=\rho_n}\mathrm{E}_0(u_n)\leq\frac{\left\|v_0\right\|^2-\left\|v_N\right\|^2}{2}+\sum_{n=1}^{N}t_n\,\mathrm{E}_0(w_n)+\left\langle v_n-v_{n-1},\ w_n\right\rangle.$$

This is almost in the desired form, however, we would like to flip the roles of $v_n/w_n$ in the final inner product term. Re-writing the right-hand side gives

$$\sum_{n=1}^{N}\left\langle v_n-v_{n-1},\ w_n\right\rangle=\left\langle v_N,\ w_N\right\rangle-\left\langle v_0,\ w_0\right\rangle+\sum_{n=1}^{N}\left\langle v_{n-1},\ w_{n-1}-w_n\right\rangle.$$

Noting $v_0=u_0$, combining the previous two equations proves the statement of Theorem C.1.2.

□

The following lemma is used to produce a sharper estimate on sequences $t_n$.

**Lemma C.1.3.** *If* $\rho_n = t_n^2 - t_{n+1}^2 + t_{n+1} \geq 0$, $t_n \geq 1$ *for all* $n \in \mathbb{N}$ *then* $t_n \leq n - 1 + t_1$.

*Proof.* This is trivially true for $n = 1$. Suppose true for $n - 1$, the condition on $\rho_{n-1}$ gives

$$t_n^2 - t_n \leq t_{n-1}^2 \leq (n - 2 + t_1)^2 = (n - 1 + t_1)^2 - 2(n - 1 + t_1) + 1.$$

Assuming the contradiction, if $t_n > n - 1 + t_1$ then we get $n - 1 + t_1 < 1$ but $t_1 \geq 1$ so this becomes $n < 1$ which completes the contradiction. □

---

**Lemma C.1.4** (Lemma 4.4.5)**.** *Let* $u_n$, $v_n$ *be generated by Algorithm 4.1,* $(n_k \in \mathbb{N})_{k=0}^{\infty}$ *be a monotone increasing sequence, and define*

$$\widetilde{\mathbb{U}}^k := \mathbb{U}^{n_k}, \qquad \widetilde{w}_k \in \operatorname*{argmin}_{u \in \widetilde{\mathbb{U}}^k} \mathrm{E}(u).$$

*If*

$$\widetilde{w}_k \in \mathbb{U}^n \qquad \textit{for all} \qquad n_k \leq n < n_{k+1}, \ k \in \mathbb{N},$$

*then for all* $K \in \mathbb{N}$, $n_K \leq N < n_{K+1}$ *we have*

$$t_N^2 \, \mathrm{E}_0(u_N) + \sum_{n=1}^{N-1} \rho_n \, \mathrm{E}_0(u_n) + \frac{\|v_N - \widetilde{w}_K\|^2}{2} \leq C + \frac{\|\widetilde{w}_K\|^2}{2} + \frac{(N+1)^2 - n_K^2}{2} \, \mathrm{E}_0(\widetilde{w}_K)$$
$$+ \sum_{k=1}^{K} \frac{n_k^2 - n_{k-1}^2}{2} \, \mathrm{E}_0(\widetilde{w}_{k-1}) + \langle v_{n_k - 1}, \ \widetilde{w}_k - \widetilde{w}_{k+1} \rangle$$

*where* $C = \frac{\|u_0 - \widetilde{w}_0\|^2 - \|\widetilde{w}_0\|^2}{2}$.

*Proof.* This is just a telescoping of the right-hand side of (C.2) with the introduction of $n_k$ and simplification $w_n = \widetilde{w}_k$,

$$\tfrac{1}{2} \|w_N\|^2 + \sum_{n=1}^{N} t_n \, \mathrm{E}_0(\Pi_n w_n) + \langle v_{n-1}, \ w_{n-1} - w_n \rangle = \tfrac{1}{2} \|\widetilde{w}_K\|^2 + \sum_{n=n_K}^{N} t_n \, \mathrm{E}_0(\widetilde{w}_K)$$
$$+ \sum_{k=1}^{K} \sum_{n=n_{k-1}}^{n_k - 1} t_n \, \mathrm{E}_0(\widetilde{w}_K) + \langle v_{n_k - 1}, \ \widetilde{w}_k - \widetilde{w}_{k+1} \rangle .$$

By Lemma C.1.3, $t_n \leq n$ so we can further simplify

$$\sum_{n=a}^{b-1} t_n \leq \sum_{n=a}^{b-1} n = (b - a) \frac{b - 1 + a}{2} \leq \frac{b^2 - a^2}{2}$$

to get the required bound. □

### C.1.2 Proof for Step 3

**Lemma C.1.5** (Lemma 4.4.6)**.** *Suppose* $\mathbb{U}^n, u_n, v_n$ *and* $n_k$ *satisfy the conditions of Lemma 4.4.5 and* $\{\widetilde{\mathbb{U}}^k\}$ *forms an* $(a_U, a_{\mathrm{E}})$*-discretisation for* $\mathrm{E}$*. If either:*

- $a_U > 1$ *and* $n_k^2 \lesssim a_{\mathrm{E}}^k a_U^{2k}$,
- *or* $a_U = 1$, $\sum_{k=1}^\infty n_k^2 a_{\mathrm{E}}^{-k} < \infty$ *and* $\sum_{k=1}^\infty \|\widetilde{w}_k - \widetilde{w}_{k+1}\| < \infty$,

*then*

$$\mathrm{E}_0(u_N) \lesssim \frac{a_U^{2K}}{N^2}$$

*for all* $n_K \leq N < n_{K+1}$.

*Proof.* Inserting the assumed rates into Lemma C.1.4 gives

$$t_N^2\,\mathrm{E}_0(u_N) + \tfrac{1}{2}\|v_N - \widetilde{w}_K\|^2 \lesssim a_U^{2K} + (N+1)^2 a_{\mathrm{E}}^{-K} + \sum_{k=1}^K n_k^2 a_{\mathrm{E}}^{-k} + a_U^k \|v_{n_k-1} - \widetilde{w}_{k-1}\| + a_U^{2k}.$$

Each case now needs its own induction. When $a_U > 1$ we simplify the inequality to

$$\begin{aligned} t_N^2\,\mathrm{E}_0(u_N) + \tfrac{1}{2}\|v_N - \widetilde{w}_K\|^2 &\lesssim a_U^{2K} + a_U^{2K+2} + \sum_{k=1}^K a_U^{2k} + a_U^k \|v_{n_k-1} - \widetilde{w}_{k-1}\| \\ &\leq C_1 \left( a_U^{2K+2} + \frac{a_U^{2K+2}}{a_U^2 - 1} + \sum_{k=1}^K a_U^k \|v_{n_k-1} - \widetilde{w}_{k-1}\| \right). \end{aligned}$$

for some $C_1$ sufficiently large. Assume $\|v_{n_k-1} - \widetilde{w}_{k-1}\| \leq C_2 a_U^k$ for $k \leq K$, then for $N = n_{K+1} - 1$ we have

$$\tfrac{1}{2}\|v_{n_{K+1}-1} - \widetilde{w}_K\|^2 \leq \frac{C_1 a_U^{2K+2}}{a_U^2 - 1}\left(a_U^2 + C_2\right).$$

If $C_2$ is sufficiently large, then $\frac{C_1 a_U^{2K+2}}{a_U^2-1}\left(a_U^2 + C_2\right) \leq \frac{1}{2} C_2^2 a_U^{2K+2}$ which completes the induction. This bounds the growth of the right hand side and so for any $N < n_{K+1}$ we have $t_N^2\,\mathrm{E}_0(u_N) \lesssim a_U^{2K}$.

When $a_U = 1$, the assumptions are stronger so the induction becomes more direct. Assuming $\|v_{n_k-1} - \widetilde{w}_{k-1}\| \leq C_2$ for $k \leq K$, there exists $C_1 > 0$ such that

$$\begin{aligned} t_{N+1}^2\,\mathrm{E}_0(u_{N+1}) + \tfrac{1}{2}\|v_{N+1} - \widetilde{w}_K\|^2 &\leq C_1 + \max_{k \leq K} \|v_{n_k-1} - \widetilde{w}_{k-1}\| \\ &\leq C_1 + C_2. \end{aligned}$$

If $C_2$ is sufficiently large then $C_1 + C_2 \leq \frac{1}{4} C_2^2$ which completes the second induction. This confirms $t_N^2\,\mathrm{E}_0(u_N) \leq \frac{1}{4} C_2^2$ is bounded uniformly. □

### C.1.3 Proof for Step 4

**Lemma C.1.6** (Lemma 4.4.7)**.** *Suppose $u_n$ and $n_k$ are sequences satisfying*

$$\mathrm{E}_0(u_N) \lesssim \frac{a_U^{2K}}{N^2} \qquad \textit{where} \qquad n_K^2 \gtrsim a_\mathrm{E}^K a_U^{2K},$$

*then*

$$\mathrm{E}_0(u_N) \lesssim \frac{1}{N^{2(1-\varepsilon)}} \qquad \textit{where} \qquad \varepsilon = \frac{\log a_U^2}{\log a_\mathrm{E} + \log a_U^2}.$$

*Proof.* The proof is direct computation,

$$\log N^2 \geq \log C + K\left(\log a_\mathrm{E} + \log a_U^2\right)$$

which leads to

$$\begin{aligned} a_U^{2K} &= \exp(K \log a_U^2) \\ &\leq \exp\left(\log N^2 \frac{\log a_U^2}{\log a_\mathrm{E} + \log a_U^2} - \frac{\log C \log a_U^2}{\log a_\mathrm{E} + \log a_U^2}\right) \\ &\lesssim N^{2\varepsilon} \end{aligned}$$

as required. □

---

**Theorem C.1.7** (Theorem 4.4.8)**.** *Let $\{\widetilde{\mathbb{U}}^k \text{ s.t. } k \in \mathbb{N}\}$ be an $(a_U, a_\mathrm{E})$-discretisation for* E *and choose any $\mathbb{U}^n$ such that*

$$\widetilde{\mathbb{U}}^k = \mathbb{U}^{n_k}, \qquad \widetilde{w}_k \in \mathbb{U}^{n_k+1} \cap \ldots \cap \mathbb{U}^{n_{k+1}-1}$$

*for all $k \in \mathbb{N}$. Compute $u_n$, $v_n$ by Algorithm 4.1 and choose $\widetilde{w}_k \in \operatorname{argmin}_{u \in \widetilde{\mathbb{U}}^k} \mathrm{E}(u)$.*

*Suppose that either:*

- *$a_U > 1$ and $n_k^2 \simeq a_\mathrm{E}^k a_U^{2k}$,*
- *or $a_U = 1$, $\sum_{k=1}^\infty n_k^2 a_\mathrm{E}^{-k} < \infty$ and $\sum_{k=1}^\infty \|\widetilde{w}_k - \widetilde{w}_{k+1}\| < \infty$,*

*then*

$$\mathrm{E}_0(u_N) \lesssim \frac{1}{N^{2(1-\varepsilon)}} \qquad \textit{where} \qquad \varepsilon = \frac{\log a_U^2}{\log a_\mathrm{E} + \log a_U^2}$$

*uniformly for $N \in \mathbb{N}$.*

*Proof.* If the conditions of this theorem are satisfied, then so are Lemmas C.1.4 to C.1.6. The final result is just the conclusion of Lemma C.1.6. □

### C.1.4 Proofs for Step 5

**Theorem C.1.8** (Theorem 4.4.9). *Let $\{\mathbb{U}^n \subset \mathbb{H}\cap\mathbb{U} \text{ s.t. } n\in\mathbb{N}\}$ be a sequence of subspaces and $n_k\in\mathbb{N}$ a monotone increasing sequence such that*

$$\widetilde{\mathbb{U}}^k := \mathbb{U}^{n_k} \ni u_{n_k-1}, \qquad \widetilde{w}_k \in \left[\operatorname*{argmin}_{u\in\widetilde{\mathbb{U}}^k} \mathrm{E}(u)\right] \cap \mathbb{U}^{n_k+1}\cap\ldots\cap\mathbb{U}^{n_{k+1}-1}$$

*for all $k\in\mathbb{N}$. Compute $u_n$, $v_n$ by Algorithm 4.1.*

*Suppose there exist $a_U, a_{\mathrm{E}}\geq 1$ such that either:*

- *$a_U>1$ and $n_k^2\lesssim a_{\mathrm{E}}^k a_U^{2k}$,*
- *or $a_U=1$, $\sum_{k=1}^\infty n_k^2 a_{\mathrm{E}}^{-k}<\infty$ and $\sum_{k=1}^\infty \|\widetilde{w}_k-\widetilde{w}_{k+1}\|<\infty$*

*and both*

$$\|\widetilde{w}_k\|\lesssim a_U^k \qquad \textit{and} \qquad \mathrm{E}_0(u_{n_K-1})\lesssim a_{\mathrm{E}}^{-K}.$$

*Whenever these conditions on $n_k$ are satisfied, then*

$$\min_{n\leq N}\mathrm{E}_0(u_n)\lesssim\frac{1}{N^{2(1-\varepsilon)}} \qquad \textit{where} \qquad \varepsilon=\frac{\log a_U^2}{\log a_{\mathrm{E}}+\log a_U^2}$$

*uniformly for $N\in\mathbb{N}$.*

*Proof.* To apply Lemma C.1.5, we need

$$n_k^2\lesssim\begin{cases}(a_{\mathrm{E}}a_U^2)^k & a_U>1\\ k^{-2}a_{\mathrm{E}}^k & a_U=1\end{cases}, \qquad \|\widetilde{w}_k\|\lesssim a_U^k, \qquad \text{and } \mathrm{E}_0(\widetilde{w}_k)\lesssim a_{\mathrm{E}}^{-k}$$

and $\sum_{k=1}^\infty\|\widetilde{w}_k-\widetilde{w}_{k+1}\|<\infty$ when $a_U=1$. The only one which is not directly assumed is easily verified,

$$\mathrm{E}_0(\widetilde{w}_k)=\min_{u\in\mathbb{U}^{n_k}}\mathrm{E}_0(u)\leq\mathrm{E}_0(u_{n_k-1})\lesssim a_{\mathrm{E}}^{-k}.$$

Therefore, the result of Lemma C.1.5 gives

$$\mathrm{E}_0(u_N)\lesssim\frac{a_U^{2K}}{t_N^2}\lesssim\frac{a_U^{2K}}{N^2}.$$

In the case $a_U=1$, this is already the optimal rate and therefore sharp. If this were sharp for general $a_U$ and E, then we gain nothing by refining early (increasing $K$ for fixed $N$) however, we can at least guarantee that refining early does not lose the optimal rate.

If we fix $N^2 \lesssim (a_{\mathrm{E}} a_U^2)^k$, then

$$\min_{n\leq N} \mathrm{E}_0(u_n) \lesssim \left\{ \begin{array}{cc} a_{\mathrm{E}}^{-k} & N > n_k \\ {a_U^{2k}}/{N^2} & N \leq n_k \end{array} \right. \lesssim N^{-2} \max(a_U^{2k}, a_U^{2k}) \lesssim N^{-2(1-\varepsilon)}$$

as required. □

---

**Lemma C.1.9** (Lemma 4.4.11)**.** *Let $\{\widetilde{\mathbb{U}}^k \text{ s.t. } k \in \mathbb{N}\}$ be a sequence of subspaces with some points $u_k \in \widetilde{\mathbb{U}}^k$ and $\widetilde{w}_k \in \operatorname{argmin}_{u\in\widetilde{\mathbb{U}}^k} \mathrm{E}(u)$. Suppose that $\|\widetilde{w}_k\| \lesssim a_U^k$. Any of the following conditions are sufficient to show that $\{\widetilde{\mathbb{U}}^k\}$ is an $(a_U, a_{\mathrm{E}})$-discretisation for $\mathrm{E}$:*

1. *Small continuous gap refinement: $\mathrm{E}_0(u_k) \leq \beta a_{\mathrm{E}}^{-k}$ for all $k \in \mathbb{N}$, some $\beta > 0$.*
2. *Small discrete gap refinement: $\mathrm{E}_0(\widetilde{w}_k) \leq \beta a_{\mathrm{E}}^{-k}$ and $\mathrm{E}_0(u_k) - \mathrm{E}_0(\widetilde{w}_{k-1}) \leq \beta a_{\mathrm{E}}^{-k}$ for all $k \in \mathbb{N}$, some $\beta > 0$.*
3. *Small relative gap refinement: $\mathrm{E}_0(u_k) - \mathrm{E}_0(\widetilde{w}_{k-1}) \leq \beta\, \mathrm{E}_0(u_k)$ for all $k \in \mathbb{N}$, some $0 < \beta \leq \frac{1}{1+a_{\mathrm{E}}}$.*
4. *Small continuous gradient refinement: $|||\partial\, \mathrm{E}(u_k)|||_* \leq \beta a_{\mathrm{E}}^{-k}$ for all $k \in \mathbb{N}$, some $\beta > 0$, and sublevel sets of $\mathrm{E}$ are $|||\cdot|||$-bounded.*
5. *Small discrete gradient refinement: $\mathrm{E}_0(\widetilde{w}_k) \leq \beta a_{\mathrm{E}}^{-k}$ and $|||\Pi_k \partial\, \mathrm{E}(u_k)|||_* \leq \beta a_{\mathrm{E}}^{-k}$ for all $k \in \mathbb{N}$, some $\beta > 0$, and sublevel sets of $\mathrm{E}$ are $|||\cdot|||$-bounded. The operator $\Pi_k \colon \mathbb{H} \to \widetilde{\mathbb{U}}^k$ is the orthogonal projection.*

*Proof.* The proof is simply to justify that the conditions of Theorem C.1.8 are met for all refinement criteria described here. $\|\widetilde{w}_k\| \lesssim a_U^k$ is enforced at every step and condition (2) guarantees the back-stop condition on $n_k$.

To complete the requirements of Theorem C.1.8, we first need to show inductively that $\mathrm{E}_0(\widetilde{w}_{k-1}) \leq C a_{\mathrm{E}}^{-k}$, then it follows that $\mathrm{E}_0(u_{n_k-1}) \leq C a_{\mathrm{E}}^{-k}$. To be explicit, when $\mathrm{E}$ has bounded sublevel sets, assume that the bound for $\{u \in \mathbb{H} \text{ s.t. } \mathrm{E}(u) \leq \mathrm{E}(u_0)\}$ is $R > 0$.

For (2) and (5) the decay of $\mathrm{E}_0(\widetilde{w}_{k-1})$ is already assumed but otherwise we need to perform the formal induction. Assume $\mathrm{E}_0(\widetilde{w}_{k-1}) \leq C a_{\mathrm{E}}^{1-k}$, then for each remaining adaptive criterion:

(1) $\mathrm{E}_0(\widetilde{w}_k) \leq \mathrm{E}_0(u_k) \leq \beta a_{\mathrm{E}}^{-k}$ by definition, the induction holds if $C \geq \beta$.

(3) $\mathrm{E}_0(\widetilde{w}_k) \leq \mathrm{E}_0(u_k) \leq \frac{\beta}{1-\beta}\, \mathrm{E}_0(\widetilde{w}_{k-1})$, the induction holds if $\beta \leq \frac{1}{1+a_{\mathrm{E}}}$.

(4) $\mathrm{E}_0(\widetilde{w}_k) \leq \mathrm{E}_0(u_k) \leq \langle \partial\, \mathrm{E}(u_k),\ u_{n_k-1} - u^* \rangle \leq 2R\beta a_{\mathrm{E}}^{-k}$, the induction holds for $C \geq 2R\beta$.

In each case, with $C$ sufficiently large, the induction holds. Now we can return to the precise condition of Theorem C.1.8, $\mathrm{E}_0(u_{n_k-1}) \leq C' a_{\mathrm{E}}^{-k}$. Assume true for $k-1$. For each adaptive criterion:

(1) $\mathrm{E}_0(u_k) \leq \beta a_{\mathrm{E}}^{-k}$ is by definition.

(2) $\mathrm{E}_0(u_k) \leq \beta a_{\mathrm{E}}^{-k} + \mathrm{E}_0(\widetilde{w}_{k-1})$ requires $C' \geq \beta + a_{\mathrm{E}}C$.

(3) $\mathrm{E}_0(u_k) \leq \frac{\beta}{1-\beta}\mathrm{E}_0(\widetilde{w}_{k-1})$ requires $C' \geq \frac{\beta}{1-\beta}C$.

(4) $\mathrm{E}_0(u_k) \leq 2R\beta a_{\mathrm{E}}^{-k}$, the induction holds for large $C'$.

(5) $\mathrm{E}_0(u_k) \leq 2R\beta a_{\mathrm{E}}^{-k} + \mathrm{E}_0(\widetilde{w}_{k-1})$, the induction holds for large $C' \geq 2R\beta + C$.

This completes the requirements of Theorem C.1.8, therefore also this proof. □

## C.2 Proof of Theorem 4.5.2

The proof of Theorem 4.5.2 is the result of the following three lemmas. The first, Lemma C.2.1, is a general quantification of the equivalence between $L^q$ and $L^2$ norms on finite dimensional sub-spaces. A special case occurs when $q = 1$ because the dual norm is a supremum rather than an integral. In Lemma C.2.2, this locality is exploited by finite element spaces, which we assumed had a basis with local support. Lemma C.2.3 then performs the computations for the $a_{\mathrm{E}}$ constant depending on the smoothness properties of E.

**Lemma C.2.1.** *Suppose* $\mathbb{H} = L^2(\Omega)$ *for some compact domain* $\Omega \subset \mathbb{R}^d$ *and* $\|\cdot\|_q \lesssim |||\cdot|||$ *for some* $q \in [1, \infty]$*. Let* $\widetilde{\mathbb{U}} \subset \mathbb{U} \cap \mathbb{H}$ *be a finite dimensional subspace with orthonormal basis* $\{e_j \text{ s.t. } j = 1, \ldots, \dim(\widetilde{\mathbb{U}})\} \subset \widetilde{\mathbb{U}}$ *and orthogonal projection* $\Pi\colon \mathbb{H} \to \widetilde{\mathbb{U}}$*. If these conditions hold, then:*

- *if* $q \geq 2$*, then*
$$\|\Pi w\| \lesssim |||w|||,$$
- *otherwise, if* $e_j \in \mathbb{U}^*$*, then*
$$\|\Pi w\| \leq \sqrt{\dim(\widetilde{\mathbb{U}})}\max_j |||e_j|||_* |||w|||,$$
- *otherwise, if* $e_j \in L^\infty(\Omega)$ *and* $|\{j : e_j(\boldsymbol{x}) \neq 0\}| \leq C$ *for almost all* $\boldsymbol{x} \in \Omega$*, then*
$$\|\Pi w\| \lesssim \sqrt{C}\max_j \|e_j\|_{L^\infty} |||w|||$$

*uniformly for all* $w \in \mathbb{H}$*.*

*Proof.* The first statement for $q \geq 2$ is from Hölder's inequality combined with the fact that compact domains have finite volume,

$$\|\Pi w\|^2 \leq \|w\|^2 = \|w\|_2^2 = \left\|w^2\right\|_1 \lesssim \|w\|_q^2 \lesssim |||w|||^2.$$

The remaining statements come from the equivalence of norms on finite dimensional spaces. Note that

$$\|\Pi w\| = \frac{\langle \Pi w,\ \Pi w\rangle}{\|\Pi w\|} = \frac{\langle \Pi w,\ w\rangle}{\|\Pi w\|} \leq \frac{|||\Pi w|||_*}{\|\Pi w\|}|||w|||,$$

therefore it is sufficient to bound $\frac{|||\cdot|||_*}{\|\cdot\|}$ on the subspace $\widetilde{\mathbb{U}}$. Switching to the given basis, for $u = \sum_j r_j e_j$ we have

$$\begin{aligned}
\|u\|^2 &= \sum_{j=1}^{\dim(\widetilde{\mathbb{U}})} r_j^2 = \|\boldsymbol{r}\|_{\ell^2}^2\,,\\
|||u|||_* &\leq \sum_{j=1}^{\dim(\widetilde{\mathbb{U}})} |r_j||||e_j|||_* \leq \max_j |||e_j|||_* \, \|\boldsymbol{r}\|_{\ell^1}\,,\\
\implies \frac{|||u|||_*}{\|u\|} &\leq \max_j |||e_j|||_* \frac{\|\boldsymbol{r}\|_{\ell^1}}{\|\boldsymbol{r}\|_{\ell^2}} \leq \max_j |||e_j|||_* \sqrt{\dim(\widetilde{\mathbb{U}})}.
\end{aligned}$$

Alternatively, we can use the inequality

$$\|\Pi w\| = \frac{\langle \Pi w,\ w\rangle}{\|\Pi w\|} \leq \frac{\|\Pi w\|_\infty}{\|\Pi w\|}\|w\|_1 \lesssim \frac{\|\Pi w\|_\infty}{\|\Pi w\|}\|w\|_q \lesssim \frac{\|\Pi w\|_\infty}{\|\Pi w\|}|||w|||.$$

This simplifies the equivalence constant because for any $u = \sum r_j e_j$ and $\mu > 1$, there exists a set of points $\boldsymbol{x} \in \Omega$ with non-zero measure such that $\|u\|_\infty \leq \mu|u(\boldsymbol{x})|$. This gives

$$\begin{aligned}
\|u\|^2 &= \sum r_j^2 \geq \sum_{e_j(\boldsymbol{x})\neq 0} r_j^2,\\
\|u\|_\infty &\leq \mu|u(\boldsymbol{x})| \leq \mu \sum_{e_j(\boldsymbol{x})\neq 0} |r_j| \, \|e_j\|_\infty\,,\\
\implies \frac{\|u\|_\infty}{\|u\|} &\leq \mu \max_j \|e_j\|_\infty \sqrt{C}.
\end{aligned}$$

This inequality holds for all $\mu > 1$ therefore also for $\mu = 1$. Essentially, with an extra smoothness assumption on $e_j$, we can reduce the dimension of the problem to $\dim(\widetilde{\mathbb{U}}) = C$ and use the previous result. □

---

**Lemma C.2.2.** *Suppose $\mathbb{H} = L^2(\Omega)$ for some compact domain $\Omega \subset \mathbb{R}^d$ and $\|\cdot\|_q \lesssim |||\cdot|||$ for some $q \in [1,\infty]$. Let $(\widetilde{\mathbb{U}}^k)_k$ be a sequence of $h$-refining finite element spaces with orthogonal projections $\widetilde{\Pi}_k \colon \mathbb{H} \to \widetilde{\mathbb{U}}^k$. If $q \geq 2$,*

$$\left\|\widetilde{\Pi}_k w\right\| \lesssim |||w|||,$$

*otherwise,*

$$\left\|\widetilde{\Pi}_k w\right\| \lesssim \sqrt{\dim(\widetilde{\mathbb{U}}^0)} h^{-\frac{kd}{2}} |||w|||$$

*uniformly for all* $w \in \mathbb{H}$.

*Proof.* Most of the conditions of Lemma C.2.1 are already satisfied. Denote $C = \dim(\widetilde{\mathbb{U}^0})$ and $\{e_j \text{ s. t. } j \in [C]\}$ the standard orthonormal basis of $\widetilde{\mathbb{U}}^0$. The scaling properties of $h$-refining finite element spaces guarantee that the value of $C$ satisfies the conditions of Lemma C.2.1 and a basis of $\widetilde{\mathbb{U}}^k$ is given by

$$\left\{\boldsymbol{x} \mapsto u_j(\alpha^{i,k}\boldsymbol{x} + \boldsymbol{\beta}^{i,k}) \qquad \text{s. t.} \qquad i = 1, \ldots, |\mathbb{M}^k|,\ j = 1, \ldots, C\right\}$$

for some $\alpha^{i,k} \in \mathbb{R}^{d\times d}$, and $\boldsymbol{\beta}^{i,k} \in \mathbb{R}^d$ such that $0 < \det(\alpha^{i,k}) \lesssim h^{-kd}$.

We now compute the scaling constant in Lemma C.2.1:

$$\frac{\left\|u_j(\alpha^{i,k} \cdot + \boldsymbol{\beta}^{i,k})\right\|_\infty}{\left\|u_j(\alpha^{i,k} \cdot + \boldsymbol{\beta}^{i,k})\right\|} = \frac{\|u_j\|_\infty}{\sqrt{\int_{\omega_i^k} e_j(\alpha^{i,k}\boldsymbol{x} + \boldsymbol{\beta}^{i,k})^2 d\boldsymbol{x}}} = \frac{\|u_j\|_\infty}{\sqrt{\det(\alpha^{i,k})^{-1}}} \lesssim h^{-\frac{kd}{2}}.$$

This value is independent of $j$ and gives the desired bound as a result of Lemma C.2.1. □

---

**Lemma C.2.3.** *Let* $(\widetilde{\mathbb{U}}^k)_k$ *be a sequence of* $h$*-refining finite element spaces of order* $p$ *with* $\widetilde{w}^k = \operatorname{argmin}_{u\in\widetilde{\mathbb{U}}^k} \mathrm{E}(u)$.

1. *If* $\mathrm{E}$ *is* $|||\cdot|||$*-Lipschitz at* $u^*$*, then* $\mathrm{E}(\widetilde{w}^k) - \mathrm{E}(u^*) \lesssim h^p|||u^*|||$.

2. *If* $\mathrm{E}$ *is* $|||\cdot|||$*-smooth at* $u^*$*, then* $\mathrm{E}(\widetilde{w}^k) - \mathrm{E}(u^*) \lesssim h^{2p}|||u^*|||$.

3. *If* $\mathrm{f}$ *is* $|||\cdot|||$*-Lipschitz at* $u^*$ *and*

$$\min_{w\in\widetilde{\mathbb{U}}^k} \{|||w - u^*||| \text{ s. t. } \mathrm{g}(w) \leq \mathrm{g}(u^*)\} \lesssim \min_{w\in\widetilde{\mathbb{U}}^k} |||w - u^*|||$$

*uniformly for* $k \in \mathbb{N}$*, then* $\mathrm{E}(\widetilde{w}^k) - \mathrm{E}(u^*) \lesssim h^p|||u^*|||$.

*Proof.* Each statement is by definition, observe

$$\mathrm{E}(w) - \mathrm{E}(u^*) \leq \mathrm{Lip}(\mathrm{E})|||w - u^*|||,$$
$$\mathrm{E}(w) - \mathrm{E}(u^*) \leq \langle \partial\,\mathrm{E}(w),\ (w - u^*)\rangle = \langle \nabla\,\mathrm{E}(w) - \nabla\,\mathrm{E}(u^*),\ w - u^*\rangle \leq \mathrm{Lip}(\nabla F)|||w - u^*|||^2,$$
$$\mathrm{E}(w) - \mathrm{E}(u^*) \leq \mathrm{f}(w) - \mathrm{f}(u^*) \leq \mathrm{Lip}(\mathrm{f})|||w - u^*|||.$$

Minimising over the right-hand side over $w$ and substituting the definition of order gives the desired result. □

## C.3 Operator norms for numerical examples

**Theorem C.3.1** (Theorem 4.6.3)**.** *Suppose $\mathcal{A}\colon \mathbb{H}\to\mathbb{R}^m$ has kernels $\psi_j\in\mathbb{H}=L^2([0,1]^d)$ for $j\in[m]$.*

*Case 1: If $\psi_j(\boldsymbol{x})=\begin{cases}1 & \boldsymbol{x}\in\mathbb{X}_j\\ 0 & \text{else}\end{cases}$ for some collection $\mathbb{X}_j\subset\Omega$ such that $\mathbb{X}_i\cap\mathbb{X}_j=\emptyset$ for all $i\neq j$, then*

$$\|\mathcal{A}\|_{L^2\to\ell^2}=\max_j\sqrt{|\mathbb{X}_j|}.$$

*Case 2: If $\psi_j(\boldsymbol{x})=\cos(\boldsymbol{a}_j\boldsymbol{\cdot}\boldsymbol{x})$ for some frequencies $\boldsymbol{a}_j\in\mathbb{R}^d$ with $|\boldsymbol{a}_j|\leq A$, then*

$$\|\mathcal{A}\|_{L^2\to\ell^2}\leq\sqrt{m},\qquad |\mathcal{A}^*\boldsymbol{r}|_{C^k}\leq m^{1-\frac{1}{q}}A^k\left\|\boldsymbol{r}\right\|_q,\qquad |\mathcal{A}^*|_{\ell^2\to C^k}\leq\sqrt{m}A^k$$

*for all $\boldsymbol{r}\in\mathbb{R}^m$ and $q\in[1,\infty]$.*

*Case 3: Suppose $\psi_j(\boldsymbol{x})=(2\pi\sigma^2)^{-\frac{d}{2}}\exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}_j|^2}{2\sigma^2}\right)$ for some regular mesh $\boldsymbol{x}_j\in[0,1]^d$ and separation $\Delta$. i.e.*

$$\{\boldsymbol{x}_j\ \text{s.t.}\ j\in[m]\}=\{\boldsymbol{x}_0+(j_1\Delta,\ldots,j_d\Delta)\ \text{s.t.}\ j_i\in[\widehat{m}]\}$$

*for some $\boldsymbol{x}_0\in\mathbb{R}^d$, $\widehat{m}:=\sqrt[d]{m}$. For all $\frac{1}{q}+\frac{1}{q^*}=1$, $q\in(1,\infty]$, we get*

$$\|\mathcal{A}\|_{L^2\to\ell^2}\leq\left((4\pi\sigma^2)^{-\frac{d}{2}}\sum_{j=-2\widehat{m},\ldots,2\widehat{m}}\exp(-\tfrac{\Delta^2}{4\sigma^2}j^2)\right)^d,$$

$$|\mathcal{A}^*\boldsymbol{r}|_{C^0}\leq(2\pi\sigma^2)^{-\frac{d}{2}}\left(\sum_{\boldsymbol{j}\in J}\exp\left(-\tfrac{q^*\Delta^2}{2\sigma^2}\max(0,|\boldsymbol{j}|-\delta)^2\right)\right)^{\frac{1}{q^*}}\left\|\boldsymbol{r}\right\|_q,$$

$$|\mathcal{A}^*\boldsymbol{r}|_{C^1}\leq\frac{(2\pi\sigma^2)^{-\frac{d}{2}}}{\sigma}\frac{\Delta}{\sigma}\left(\sum_{\boldsymbol{j}\in J}(|\boldsymbol{j}|+\delta)^{q^*}\exp\left(-\tfrac{q^*\Delta^2}{2\sigma^2}\max(0,|\boldsymbol{j}|-\delta)^2\right)\right)^{\frac{1}{q^*}}\left\|\boldsymbol{r}\right\|_q,$$

$$|\mathcal{A}^*\boldsymbol{r}|_{C^2}\leq\frac{(2\pi\sigma^2)^{-\frac{d}{2}}}{\sigma^2}\left(\sum_{\boldsymbol{j}\in J}\left(1+\tfrac{\Delta^2}{\sigma^2}(|\boldsymbol{j}|+\delta)^2\right)^{q^*}\exp\left(-\tfrac{q^*\Delta^2}{2\sigma^2}\max(0,|\boldsymbol{j}|-\delta)^2\right)\right)^{\frac{1}{q^*}}\left\|\boldsymbol{r}\right\|_q,$$

*where $\delta=\frac{\sqrt{d}}{2}$ and $J=\{\boldsymbol{j}\in\mathbb{Z}^d\ \text{s.t.}\ \|\boldsymbol{j}\|_{\ell^\infty}\leq 2\widehat{m}\}$. The case for $q=1$ can be inferred from the standard limit of $\|\cdot\|_{q^*}\to\|\cdot\|_\infty$ for $q^*\to\infty$. For $\Delta\ll\sigma$ (i.e. high resolution data), we get the scaling behaviour*

$$\|\mathcal{A}\|_{L^2\to\ell^2}\lesssim\Delta^{-d},\qquad |\mathcal{A}^*\boldsymbol{r}|_{C^k}\lesssim\sigma^{-k}\Delta^{-\frac{d}{q^*}}\left\|\boldsymbol{r}\right\|_q,\qquad |\mathcal{A}^*|_{\ell^2\to C^k}\lesssim\sigma^{-k}\Delta^{-\frac{d}{2}},$$

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

*Proof case 1:* From Lemma 4.6.1 we have

$$(\mathcal{A}\mathcal{A}^*)_{i,j} = \left\langle \mathbb{1}_{\mathbb{X}_i},\ \mathbb{1}_{\mathbb{X}_j} \right\rangle = |\mathbb{X}_i \cap \mathbb{X}_j| = \left\{ \begin{array}{cc} |\mathbb{X}_i| & i = j \\ 0 & i \neq j \end{array} \right. .$$

Therefore, $\mathcal{A}\mathcal{A}^*$ is a diagonal matrix and $\|\mathcal{A}\mathcal{A}^*\|_{\ell^2\to\ell^2} = \max_j |\mathbb{X}_j|$ completes the result. □

---

*Proof case 2:* $\psi_j$ are not necessarily orthogonal however $|\langle\psi_i,\ \psi_j\rangle| \leq 1$ therefore we can estimate

$$\|\mathcal{A}\mathcal{A}^*\|_{\ell^2\to\ell^2} \leq \|\mathcal{A}\mathcal{A}^*\|_{\ell^\infty\to\ell^\infty} \leq m.$$

Now looking to apply Lemma 4.6.2, note $\left\|\nabla^k\psi_j\right\|_\infty \leq A^k$, therefore

$$\begin{aligned} |\mathcal{A}^*\boldsymbol{r}|_{C^k} &\leq A^k m^{\frac{1}{q^*}} \|\boldsymbol{r}\|_q = A^k m^{1-\frac{1}{q}} \|\boldsymbol{r}\|_q, \\ |\mathcal{A}^*|_{\ell^2\to C^k} &\leq A^k \min_{q\in[1,\infty]} m^{1-\frac{1}{q}} \sqrt{m}^{\max(0,2-q)} = \sqrt{m}A^k. \end{aligned}$$

□

---

*Proof of asymptotic case 3:* In the Gaussian case, we build our approximations around the idea that sums of Gaussians on a regular grid look like a discretised integral. The first example can be used to approximate the operator norm. Computing the inner products gives

$$\begin{aligned} \langle\psi_i,\ \psi_j\rangle = (2\pi\sigma^2)^{-d}\int_{[0,1]^d} \exp\left(-\tfrac{|\boldsymbol{x}-\boldsymbol{x}_i|^2}{2\sigma^2} - \tfrac{|\boldsymbol{x}-\boldsymbol{x}_j|^2}{2\sigma^2}\right) &\leq (2\pi\sigma^2)^{-d}(\pi\sigma^2)^{\frac{d}{2}} \exp\left(-\tfrac{|\boldsymbol{x}_i-\boldsymbol{x}_j|^2}{4\sigma^2}\right) \\ &= (4\pi\sigma^2)^{-\frac{d}{2}} \exp\left(-\tfrac{|\boldsymbol{x}_i-\boldsymbol{x}_j|^2}{4\sigma^2}\right). \end{aligned}$$

Estimating the operator norm,

$$\begin{aligned} \|\mathcal{A}\mathcal{A}^*\|_{\ell^2\to\ell^2} \leq \|\mathcal{A}\mathcal{A}^*\|_{\ell^\infty\to\ell^\infty} &= \max_{i\in[m]}\sum_{j=1}^m |\langle\psi_i,\ \psi_j\rangle| \\ &= \max_{i\in[m]}(4\pi\sigma^2)^{-\frac{d}{2}} \sum_{j_1,\ldots,j_d\in[\widehat{m}]} \exp\left(-\frac{(j_1\Delta - i_1\Delta)^2 + \ldots + (j_d\Delta - i_d\Delta)^2}{4\sigma^2}\right) \\ &\leq (4\pi\sigma^2)^{-\frac{d}{2}} \sum_{\boldsymbol{j}\in\mathbb{Z}^d\cap[-\widehat{m},\widehat{m}]^d} \exp\left(-\frac{(j_1\Delta)^2 + \ldots + (j_d\Delta)^2}{4\sigma^2}\right) \\ &= \frac{(4\pi)^{-\frac{d}{2}}}{\Delta^d} \sum_{\boldsymbol{j}\in\mathbb{Z}^d\cap[-\widehat{m},\widehat{m}]^d} \exp\left(-\frac{1}{4}\left|\boldsymbol{j}\frac{\Delta}{\sigma}\right|^2\right)\frac{\Delta^d}{\sigma^d} \\ &\sim \frac{(4\pi\sigma^2)^{-\frac{d}{2}}}{\Delta^d}\int_{\mathbb{R}^d} \exp\left(-\tfrac{|\boldsymbol{x}|^2}{4\sigma^2}\right) \quad = \Delta^{-d}. \end{aligned}$$

This is a nice approximation because it depends only on $\Delta$, overcoming the $\frac{1}{\sigma}$ scaling. To convert this into an upper bound, we just compute the sum explicitly. In particular, as the sum factorises over dimensions,

$$\|\mathcal{A}\mathcal{A}^*\|_{\ell^2\to\ell^2} \le (4\pi\sigma^2)^{-\frac{d}{2}}\left(\sum_{j=-\widehat{m},\dots,\widehat{m}} \exp\left(-\tfrac{\Delta^2}{4\sigma^2}j^2\right)\right)^d .$$

Applying the same ideas to Lemma 4.6.2, note

$$\begin{aligned}
|\psi_j(\boldsymbol{x})| &= |\psi_j(\boldsymbol{x})| && = (2\pi\sigma^2)^{-\frac{d}{2}}\exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}_j|^2}{2\sigma^2}\right),\\
|\nabla\psi_j(\boldsymbol{x})| &= \left|\frac{\boldsymbol{x}-\boldsymbol{x}_j}{\sigma^2}\psi_j(\boldsymbol{x})\right| && = \frac{(2\pi\sigma^2)^{-\frac{d}{2}}}{\sigma}\frac{|\boldsymbol{x}-\boldsymbol{x}_j|}{\sigma}\exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}_j|^2}{2\sigma^2}\right),\\
|\nabla^2\psi_j(\boldsymbol{x})| &= \left|\frac{1}{\sigma^2}+\frac{(\boldsymbol{x}-\boldsymbol{x}_j)(\boldsymbol{x}-\boldsymbol{x}_j)^\top}{\sigma^4}\right|\psi_j(\boldsymbol{x}) && = \frac{(2\pi\sigma^2)^{-\frac{d}{2}}}{\sigma^2}\left(1+\frac{|\boldsymbol{x}-\boldsymbol{x}_j|^2}{\sigma^2}\right)\exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}_j|^2}{2\sigma^2}\right).
\end{aligned}$$

With the substitution $\boldsymbol{x}=\frac{\Delta}{\sigma}\boldsymbol{j}$, the asymptotic bounds on these are clear:

$$\begin{aligned}
\sum_{\boldsymbol{j}\in[\widehat{m}]^d}|\psi_{\boldsymbol{j}}(\boldsymbol{x})|^{q^*} &\lesssim (2\pi\sigma^2)^{-\frac{dq^*}{2}}\frac{\sigma^d}{\Delta^d}\int_{\mathbb{R}^d}\exp\left(-\frac{q^*|\boldsymbol{x}|^2}{2}\right),\\
\sum_{\boldsymbol{j}\in[\widehat{m}]^d}|\nabla\psi_{\boldsymbol{j}}(\boldsymbol{x})|^{q^*} &\lesssim \frac{(2\pi\sigma^2)^{-\frac{dq^*}{2}}}{\sigma}\frac{\sigma^d}{\Delta^d}\int_{\mathbb{R}^d}|\boldsymbol{x}|^{q^*}\exp\left(-\frac{q^*|\boldsymbol{x}|^2}{2}\right),\\
\sum_{\boldsymbol{j}\in[\widehat{m}]^d}|\nabla^2\psi_{\boldsymbol{j}}(\boldsymbol{x})|^{q^*} &\lesssim \frac{(2\pi\sigma^2)^{-\frac{dq^*}{2}}}{\sigma^2}\frac{\sigma^d}{\Delta^d}\int_{\mathbb{R}^d}(1+|\boldsymbol{x}|^2)^{q^*}\exp\left(-\frac{q^*|\boldsymbol{x}|^2}{2}\right).
\end{aligned}$$

□

---

*Proof of precise case 3:* In the asymptotic case we have shown

$$\begin{aligned}
|\psi_j(\boldsymbol{x})| &= |\psi_j(\boldsymbol{x})| && = (2\pi\sigma^2)^{-\frac{d}{2}}\exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}_j|^2}{2\sigma^2}\right),\\
|\nabla\psi_j(\boldsymbol{x})| &= \left|\frac{\boldsymbol{x}-\boldsymbol{x}_j}{\sigma^2}\psi_j(\boldsymbol{x})\right| && = \frac{(2\pi\sigma^2)^{-\frac{d}{2}}}{\sigma}\frac{|\boldsymbol{x}-\boldsymbol{x}_j|}{\sigma}\exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}_j|^2}{2\sigma^2}\right),\\
|\nabla^2\psi_j(\boldsymbol{x})| &= \left|\frac{1}{\sigma^2}+\frac{(\boldsymbol{x}-\boldsymbol{x}_j)(\boldsymbol{x}-\boldsymbol{x}_j)^\top}{\sigma^4}\right|\psi_j(\boldsymbol{x}) && = \frac{(2\pi\sigma^2)^{-\frac{d}{2}}}{\sigma^2}\left(1+\frac{|\boldsymbol{x}-\boldsymbol{x}_j|^2}{\sigma^2}\right)\exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}_j|^2}{2\sigma^2}\right).
\end{aligned}$$

We now wish to sum over $j=1,\dots,m$ and produce an upper bound on these, independent of $t$. To do so we will use the following lemma.

**Lemma C.3.2.** *Suppose $q > 0$. If the polynomial $p(|\boldsymbol{x}|) = \sum p_k |\boldsymbol{x}|^k$ has non-negative coefficients and $\boldsymbol{x} \in [-m, m]^d$, then*

$$\sum_{\substack{\boldsymbol{j} \in \mathbb{Z}^d \\ \|\boldsymbol{j}\|_{\ell^\infty} \leq m}} p(|\boldsymbol{j} - \boldsymbol{x}|) \exp\left(-\tfrac{q|\boldsymbol{j}-\boldsymbol{x}|^2}{2}\right) \leq \left[\sum_{\substack{\boldsymbol{j} \in \mathbb{Z}^d \\ \|\boldsymbol{j}\|_{\ell^\infty} \leq 2m}} p(|\boldsymbol{j}| + \delta) \exp\left(-\frac{q \max(0, |\boldsymbol{j}| - \delta)^2}{2}\right)\right]$$

*where $\delta := \frac{\sqrt{d}}{2}$.*

*Proof.* There exists $\widehat{\boldsymbol{x}} \in [-\frac{1}{2}, \frac{1}{2}]^d$ such that $\boldsymbol{x} + \widehat{\boldsymbol{x}} \in \mathbb{Z}^d$, therefore

$$\begin{aligned}
\sum_{\substack{\boldsymbol{j} \in \mathbb{Z}^d \\ \|\boldsymbol{j}\|_{\ell^\infty} \leq m}} p(|\boldsymbol{j} - \boldsymbol{x}|) \exp\left(-\tfrac{q|\boldsymbol{j}-\boldsymbol{x}|^2}{2}\right) &= \sum_{\substack{\boldsymbol{j} \in \mathbb{Z}^d \\ \|\boldsymbol{j}\|_{\ell^\infty} \leq m}} p(|\boldsymbol{j} - (\boldsymbol{x} + \widehat{\boldsymbol{x}}) + \widehat{\boldsymbol{x}}|) \exp\left(-\tfrac{q|\boldsymbol{j}-(\boldsymbol{x}+\widehat{\boldsymbol{x}})+\widehat{\boldsymbol{x}}|^2}{2}\right) \\
&\leq \sum_{\substack{\boldsymbol{j} \in \mathbb{Z}^d \\ \|\boldsymbol{j}\|_{\ell^\infty} \leq 2m}} p(|\boldsymbol{j} + \widehat{\boldsymbol{x}}|) \exp\left(-\tfrac{q|\boldsymbol{j}+\widehat{\boldsymbol{x}}|^2}{2}\right) \\
&\leq \sum_{\substack{\boldsymbol{j} \in \mathbb{Z}^d \\ \|\boldsymbol{j}\|_{\ell^\infty} \leq 2m}} p(|\boldsymbol{j}| + \delta) \exp\left(-\tfrac{q \max(0, |\boldsymbol{j}| - \delta)^2}{2}\right)
\end{aligned}$$

as $|\widehat{\boldsymbol{x}}| \leq \delta$ and $p$ has non-negative coefficients. □

Now, continuing the proof of Theorem C.3.1, for $\widehat{m} = \sqrt[d]{m}$, $\delta = \frac{\sqrt{d}}{2}$ and $J = \{\boldsymbol{j} \in \mathbb{Z}^d \text{ s. t. } \|\boldsymbol{j}\|_{\ell^\infty} \leq 2\widehat{m}\}$, Lemma C.3.2 bounds

$$\begin{aligned}
\sum_{j=1}^m |\psi_j(\boldsymbol{x})|^{q^*} &\leq (2\pi\sigma^2)^{-\frac{dq^*}{2}} \left[\sum_{j \in J} \exp\left(-\frac{q^*\Delta^2}{2\sigma^2} \max(0, |\boldsymbol{j}| - \delta)^2\right)\right] \\
\sum_{j=1}^m |\nabla\psi_j(\boldsymbol{x})|^{q^*} &\leq \frac{(2\pi\sigma^2)^{-\frac{dq^*}{2}}}{\sigma^{q^*}} \frac{\Delta^{q^*}}{\sigma^{q^*}} \left[\sum_{j \in J} (|\boldsymbol{j}| + \delta)^{q^*} \exp\left(-\frac{q^*\Delta^2}{2\sigma^2} \max(0, |\boldsymbol{j}| - \delta)^2\right)\right] \\
\sum_{j=1}^m |\nabla^2\psi_j(\boldsymbol{x})|^{q^*} &\leq \frac{(2\pi\sigma^2)^{-\frac{dq^*}{2}}}{\sigma^{2q^*}} \left[\sum_{j \in J} \left(1 + \frac{\Delta^2}{\sigma^2}(|\boldsymbol{j}| + \delta)^2\right)^{q^*} \exp\left(-\frac{q^*\Delta^2}{2\sigma^2} \max(0, |\boldsymbol{j}| - \delta)^2\right)\right]
\end{aligned}$$

for all $\boldsymbol{x} \in \Omega$. In a worst case, this is $O(2^d m)$ time complexity however the summands all decay faster than exponentially and so should converge very quickly. □