Sparsity-based algorithms for inverse problems

Ganguly, P.S.

Chapter 1

Introduction

In this chapter we give a general introduction to the field of inverse problems and algorithmic approaches to solving such problems in a few application areas. We also introduce the reader to notions of sparsity and show how sparsity is used to tackle the various research questions investigated in this thesis.

1.1 Background

Scientific questions can be broadly divided into two kinds. The first kind seeks to question the effects of a set of causal factors while the second seeks to determine the causal factors given the effects. In this thesis we deal with the latter type of questions.

We shall restrict ourselves to situations where the effects are observations or measurements of a physical system, and the causal factors are certain variables that characterize the system. One common starting point in this case is to construct a simplified representation or physical model of the system.

An example of this process of model building is the physical theory of the interaction of light with matter. Such a theory enables us to calculate, among other things, the interaction of X-rays passing through a three-dimensional object. The measurements from this system are images – two-dimensional snapshots of the X-ray beam after it emerges from the object. These snapshots can be obtained using an X-ray detection system and compared against our prediction from the physical theory. It turns out that in this case our predictions match the experimental measurements well, thus validating the correctness of our theory.

The problem described above is a *direct* or *forward* problem, where we predict the effects given causes and a reliable model. Complementary to this problem is the *inverse problem*, where we want to infer the physical properties of the 3D object, in particular its capacity to interact with X rays, using a set of 2D images. An illustration of both problems is shown in Figure 1.1. It turns out that the inverse problem of reconstruction brings about a different set of challenges to the forward problem of projection, and in order to solve the former problem we must make further assumptions about the 3D object.

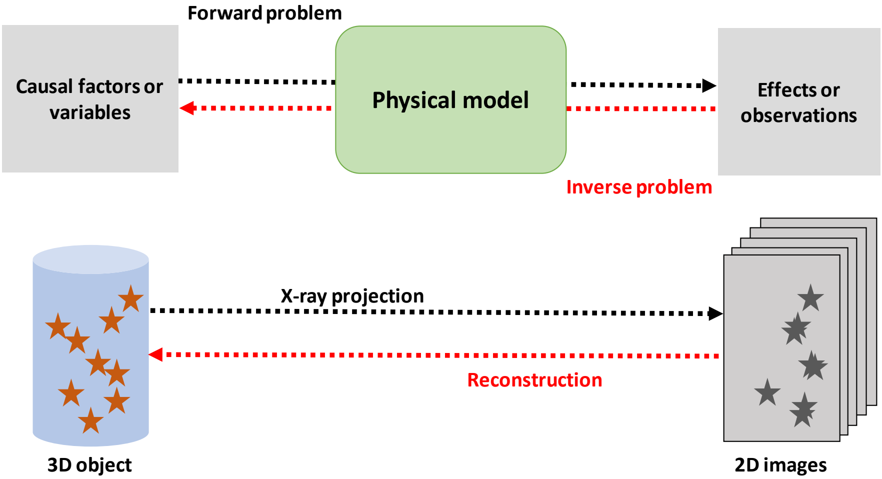


Figure 1.1: A forward problem is one where we predict the effects of a set of causal factors given a physical model of the system. An inverse problem is one where we invert this process. An example of a forward problem is the calculation of a 2D X-ray projection of a 3D object; the inversion of a set of such images to infer the 3D object is the inverse problem of reconstruction.

One such assumption that is central to the work in this thesis comes from the notion of sparsity. Broadly speaking, sparsity is the assumption that only a small set of variables or causal factors is sufficient to explain the measurements of a system. In the example of X-ray reconstruction above, sparsity could mean that the object is made of a small set of discrete constituents. Using this as prior knowledge of our 3D object makes the inversion procedure much more reliable. Stated differently, assuming sparsity enables us to limit our search for causal factors to a small set.

In this thesis we study several application areas where notions of sparsity yield practical algorithms for inverse problems. In the next section we first introduce these application areas and discuss the forward problems therein. Next we present the mathematical framework of inverse problems and discuss ways to include sparsity in this framework. In the penultimate section, we return to our application areas and show how practical algorithms can be designed to tackle sparse inverse problems in these areas. Finally, we present four research questions that are investigated in the following chapters of this thesis, and provide a brief abstract of our methods and contributions.

1.2 Application areas

In this section we introduce three application areas that were studied in this thesis, and give some mathematical background to these areas.

1.2.1 X-ray computed tomography

X-ray computed tomography (CT) is a powerful method to visualize and obtain quantitative information about the inside of an object non-destructively. X-ray CT is widely used in medical settings for diagnostic purposes [1], in materials science for studying structural changes in materials [2] and in cultural heritage for probing the construction of art objects [3].

In this imaging modality, an X-ray beam is used to generate projection images of an object of interest. The flux of the X-ray beam changes as it passes through the object according to the Beer–Lambert law:

$$I = I_0 \, e^{-\int_0^l \mu(z) \, dz} \,, \tag{1.1}$$

where I_0 is the flux of the incident X-ray beam, I is the flux after the beam has passed through a distance l inside the object and μ is the attenuation coefficient that denotes the capacity of the materials in the object to absorb X rays. Dividing both sides of (1.1) by I_0 and taking the logarithm, we arrive at the linear projection model of X-ray CT:

$$\log\left(\frac{I}{I_0}\right) = -\int_0^l \mu(z) \, dz \tag{1.2}$$

Many different experimental setups are used for X-ray CT depending on the application, but in most setups the basic acquisition strategy consists of rotating the sample with respect to the incident X-ray beam to acquire measurements along several projection angles. The emergent X-ray beam after absorption by the sample is detected using a detection system. In this thesis we focus on *parallel-beam* CT, where the distance between X-ray source and object is large enough to approximate the incident rays as being parallel to each other. This is the setup shown in Figure 1.2. Using (1.2) the forward projection of a 2D object $f(x, y)$ taken along a projection angle θ, $P_\theta(t)$, can then be written as

$$P_\theta(t) := \mathcal{R}(f) = -\iint_{\mathbb{R}^2} f(x, y) \, \delta(x \cos\theta + y \sin\theta - t) \, dx \, dy \,, \tag{1.3}$$

where the function $f : \mathbb{R}^2 \to \mathbb{R}$ is a finite integrable function with bounded support describing the attenuation of the object. Note that $\mu(z)$ in (1.2) is equivalent to the function $f(x, y)$ evaluated on the line given by $x\cos\theta + y\sin\theta = t$. $\mathcal{R}(f)$ is known as the Radon transform of the function $f(x, y)$, and δ denotes the delta function. Using (1.3) a set of projections $P_\theta(t)$, $\theta \in [0, \pi)$ can be acquired and rearranged to give a *sinogram*. In Figure 1.2, we show a popular analytical object – known as the Shepp-Logan phantom – along with its sinogram.

The tomographic reconstruction problem refers to the inversion of (1.3) to yield a suitable function $f(x, y)$ from a set of measurements $P_\theta(t)$, $\theta \in [0, \pi)$. In Section 1.4 we shall return to this inverse problem and discuss it in more detail.

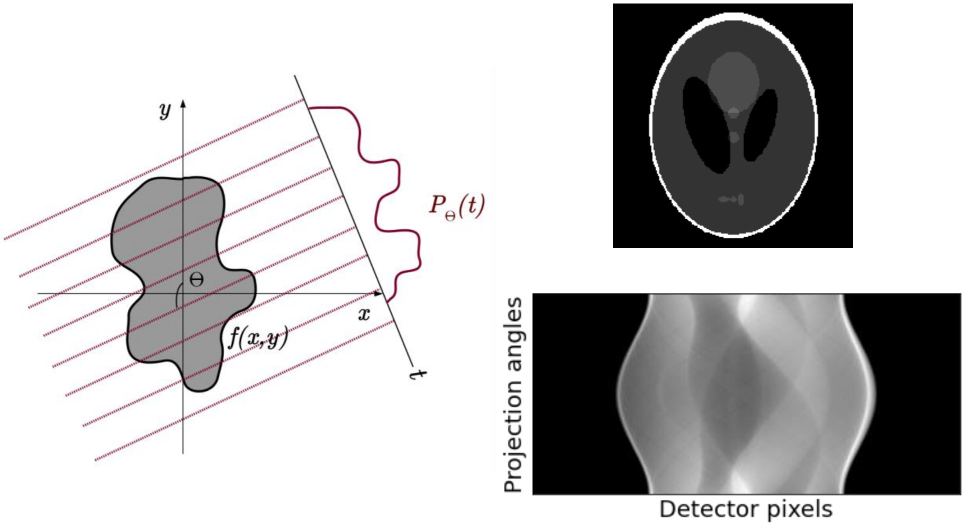


Figure 1.2: Parallel-beam X-ray CT (left), Shepp–Logan phantom (top right) and its sinogram (bottom right).

1.2.2 Electron tomography

The next application area of relevance to this thesis is electron tomography (ET). ET is the method of choice for imaging nanoparticles and biological macromolecules at atomic or near-atomic resolutions. Imaging with an electron beam allows for much higher resolutions compared to X-ray imaging because of the shorter wavelength of electrons [4], [5].

Images in ET are generated by passing a focused electron beam through a sample. In one common modality, each projection image is generated in transmission electron microscopy (TEM) mode. In this mode, the whole sample is irradiated with the incident electron beam at the same time and the transmitted electron beam is detected. Alternatively, the electron beam can be focused to scan the sample one small area at a time. This mode is known as scanning transmission electron microscopy (STEM). To obtain images at different projection angles, the sample is tilted with respect to the beam. For thin samples, the linear projection model (1.2) holds for each image in the tilt series. Similar to X-ray CT, the inverse problem in ET consists of inverting the forward model to infer the structure of nanoparticles and macromolecules from their projections.

In this thesis, we present methods for two different applications of ET. These are atomic-resolution ET and cryoET. In Section 1.4, we focus on each of these areas separately and state the inverse problems we investigated for each.

1.2.3 Vascular network formation

The final application area we study in this thesis is of relevance to developmental and cancer biology. Vasculogenesis is the process by which a primitive circulatory system is generated in vertebrates. Following the generation of a primitive network, new blood vessels arise by sprouting and expanding, in a process known as angiogenesis. Angiogenesis also occurs in certain types of cancer, where it contributes to tumour maintenance and metastasis.

Understanding how individual cells organize to form mature vascular networks is a long-standing question. In particular, the contribution of cell–cell interactions and environmental cues are still a topic of research. One way to investigate the conditions for vascular network formation is using computer simulations, where different cell–cell interactions and environmental cues can be prescribed and the resulting long-term dynamics can be studied. Simulation studies are particularly effective because all the parameters of a chosen model can be adjusted and different parameter regimes, which may not be easy to probe in experimental studies, are easily simulated.

Different simulation paradigms have been used in the literature to simulate vascular network generation. One paradigm is cellular Potts model, a lattice-based simulation where cells are represented as patches of interacting spins. A complementary paradigm is a lattice-free particle-based model, where each cell is represented by an ellipse and is assumed to interact with all other cells in a prescribed neighbourhood. The forward model of this particle-based paradigm is given by a Langevin equation:

$$\frac{d\boldsymbol{v}_i}{dt} = \frac{1}{m_i}\Big(-\tau \boldsymbol{v}_i + \sum_{j \neq i} \frac{\boldsymbol{x}_i - \boldsymbol{x}_j}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|} F_{ij} + \boldsymbol{\eta} \Big), \quad \boldsymbol{v}_i = \frac{d\boldsymbol{x}_i}{dt}, \tag{1.4}$$

where $\boldsymbol{x}_i$ denotes the position of cell i at time t, $\boldsymbol{v}_i$ is the velocity of cell i, m_i is the mass of cell i, τ is the so-called damping constant, F_{ij} is the pairwise interaction between cells i and j, and $\boldsymbol{\eta}$ is a stochastic noise term.

Using the above equation, the long-time dynamics of cells can be simulated for different parameter values. The steady-state solutions can then be used to determine the parameter regions – and hence conditions – for network formation.

An alternative approach is to start directly from experimental observations of network formation and infer the interaction terms and parameter values in (1.4). In Section 1.4, we shall return to this inverse problem and describe our methods to tackle it.

1.3 Inverse problems

In the previous section, we described several application areas and the forward problems that arise in each of them. We also mentioned briefly the inverse to these forward problems. In this section we discuss the mathematical framework of inverse problems and describe ways to solve such problems reliably.

Inverse problems arise in many areas of science and engineering, where the goal is to infer specific variables given measurements of a system and a reliable physical model [6].

Mathematically, this translates to inferring x from data y, where the two are related by the following equation:

$$A(x) = y, \tag{1.5}$$

where A is the forward model.

One of the examples described above is that of tomographic reconstruction. The forward problem of tomography is the projection of a 3D object along a set of projection angles, while the inverse is combining the information from a set of projections to obtain a reconstruction. In this example, the forward problem has a well-defined solution but the inverse problem does not.

One way to understand the difference between forward and inverse problems is the notion of *well-posedness*. A mathematical problem is said to be well-posed if its solution exists for arbitrary data (existence) and is unique (uniqueness). Additionally, the solution must depend continuously on the data such that small changes in the data result in correspondingly small changes in the solution (stability). Problems that do not meet these conditions are known as *ill-posed*.

Some physical intuition regarding the ill-posedness of inverse problems is obtained by using the idea of entropy from the second law of thermodynamics and information theory. Forward problems are generally those that describe physical phenomena oriented along the cause–effect sequence [6]. The cause–effect sequence is determined by the second law of thermodynamics, which posits an increase in total entropy in the direction of time evolution. This means that the solution to a direct problem has lower "information content" than the data. The opposite is true for an inverse problem, where data with lower information content must be used to infer unknowns with higher information content.

The ill-posedness of inverse problems – specifically the fact that small variations in the data (caused, for e.g. , by measurement noise) lead to large variations in the solution – makes it difficult to obtain a meaningful solution to an inverse problem. This is addressed by using prior knowledge about the physical system being studied. The mathematical theory that deals with this is called regularization.

An illustration of regularization is provided by the use of Tikhonov regularization in X-ray CT. The discrete formulation of the X-ray CT problem is given by:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}, \tag{1.6}$$

where $\boldsymbol{A}$ is the linear forward operator which amounts to the discretized version of the Radon transform (1.3), $\boldsymbol{y}$ is a vector of discrete projection data and $\boldsymbol{x}$ is the unknown discretized reconstruction. The reconstruction problem can then be stated as an optimization problem where we seek to minimize the difference with respect to projection data. The least-squares solution to the discrete reconstruction problem is

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \tag{1.7}$$

An example reconstruction of the Shepp-Logan phantom using a least-squares strategy is shown in Figure 1.3. A common way to regularize this problem is to minimize not just the discrepancy with respect to the projection data but also the energy of the solution, defined

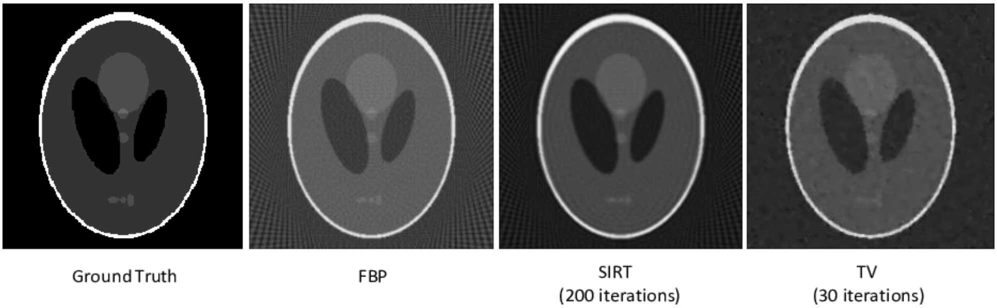


Figure 1.3: Tomographic reconstructions of the Shepp–Logan phantom using filtered back-projection (FBP), the simultaneous iterative reconstruction technique (SIRT) and total variation (TV) minimization. SIRT solves a least-squares problem with added preconditioning; TV solves a regularized least-squares problem that penalizes large gradients in the reconstructed image. All reconstructions were performed using the Astra Toolbox [7] and the Operator Discretization Library (ODL) [8].

as its ℓ^2-norm. This regularization, known as Tikhonov regularization, then amounts to the optimization problem:

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad \|\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2, \tag{1.8}$$

where $\lambda > 0$ – the regularization parameter – adjusts the relative weighting of the two terms in the optimization objective.

1.3.1 Sparse inverse problems

The example of regularization shown above penalizes the energy of the solution. In the last three decades, a different form of prior knowledge has been shown to be a powerful technique for solving a host of inverse problems [9], [10]. This prior knowledge relates to the *sparsity* of the unknown vector $\boldsymbol{x}$. One notion of sparsity is given by the number of nonzero elements of the vector $\boldsymbol{x}$, which is called the ℓ^0 "norm". The ℓ^0 "norm" of a vector $\boldsymbol{x} \in \mathbb{R}^d$ is given by

$$|\boldsymbol{x}|_0 := \sum_{i=1}^{d} |x_i|^0, \tag{1.9}$$

and the vector $\boldsymbol{x}$ is said to be s-sparse if

$$|\boldsymbol{x}|_0 \leq s. \tag{1.10}$$

A sparse inverse problem is one where we look for the sparsest solution that explains the observed data. Mathematically, we can state a sparse inverse problem as a constrained optimization problem where the goal is to

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad |\boldsymbol{x}|_0 \qquad \text{subject to} \quad \boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{y}. \tag{1.11}$$

In some scenarios a reformulation of the optimization problem may be more appropriate. For example, we may choose to relax the exact equality in the constraint to account for measurement noise

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad |\boldsymbol{x}|_0 \qquad \text{subject to} \quad \|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{y}\|_2^2 \leq \epsilon. \tag{1.12}$$

Or, we could switch the objective function with the constraint:

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad \|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{y}\|_2^2 \qquad \text{subject to} \quad |\boldsymbol{x}|_0 \leq s. \tag{1.13}$$

Objective and constraint functions may also be added to result in an optimization problem analogous to Tikhonov regularization (1.8):

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad \|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda|\boldsymbol{x}|_0. \tag{1.14}$$

The ℓ^0 term in the above formulations makes the optimization problem nonconvex, and thus sensitive to initialization. A convex surrogate is achieved by replacing the ℓ^0 term with the ℓ^1 norm of $\boldsymbol{x}$. The convex surrogate of (1.14) is

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad \|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{x}\|_1. \tag{1.15}$$

The convex formulation above can be solved with guarantees on existence and convergence of the solution. However, nonconvex optimization methods, such as greedy pursuit and simulated annealing, have also been used to solve the ℓ^0 minimization problem directly [11].

In many applications, tomographic reconstruction being one of them, the unknown reconstruction is not sparse per se, but can be sparsely coded in a suitable orthonormal basis. For e.g., images can be assumed to be piecewise constant, which implies sparsity in the space of gradients. This results in the total variation (TV) regularization method that results in surprisingly good reconstructions even for heavily undersampled data:

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad \|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda\||\nabla\boldsymbol{x}|\|_1. \tag{1.16}$$

Reconstruction of the Shepp–Logan phantom using TV regularization is shown in Figure 1.3.

In some applications like in atomic-resolution ET, discretization of the reconstruction $\boldsymbol{x}$ as an image may not be the most suitable. In situations where $\boldsymbol{x}$ is a set of unknown cardinality (e.g. the atomic coordinates in a nanoparticle, with an unknown number of atoms), the optimization problem (1.15) is no longer convex in the space of atomic coordinates. A convex formulation in such situations is achieved by lifting the problem to the vector space of measures [12].

1.4 Computational solution of inverse problems using sparsity

In this section, we return to the application areas introduced in Section 1.2 and present ways to include sparsity in the design of efficient algorithms. At the end of each subsection, we summarise our use of sparsity in that particular problem in a red box.

1.4.1 Sparse design of implementation-adapted filters for X-ray CT

The discretized tomographic reconstruction problem is the estimation of $\boldsymbol{x}$ from equation (1.6). There exist several methods to obtain an estimate of the reconstruction $\boldsymbol{x}$ given data $\boldsymbol{y}$. Direct methods that use discretized inversions of the continuous Radon transform (1.3) are some of the most popular methods due to their speed. An exact inversion of the continuous Radon transform is possible and leads to the following inversion formula for the function $f(x, y)$:

$$f(x,y) = \int_0^\pi q_\theta(x\cos\theta + y\sin\theta)\, d\theta\,, \quad q_\theta(t) = \int_\infty^{-\infty} |u|\hat{P}_\theta(u)e^{2\pi i u t}\, du, \tag{1.17}$$

where $|u|$ is known as the ramp filter in Fourier space and $\hat{P}_\theta$ denotes projection data in Fourier space. Filtered backprojection (FBP), a real-space direct method, computes a discretized version of the above inversion formula, such that

$$f(x,y) \approx \sum_{\theta\in\Theta}\sum_{\tau\in T} P_\theta(\tau)h_\theta(x\cos\theta + y\sin\theta - \tau),$$

where h is a discretized filter in real space. Equivalently, starting from the algebraic equation (1.6), the FBP reconstruction $\tilde{\boldsymbol{x}}_{\text{FBP}}$ of projection data $\boldsymbol{y}$ is given by

$$\tilde{\boldsymbol{x}}_{\text{FBP}} = \boldsymbol{A}^T\left(\boldsymbol{y} * \boldsymbol{h}\right) = \boldsymbol{A}^T\boldsymbol{C}_h\,\boldsymbol{y}, \tag{1.18}$$

where $\boldsymbol{C}_h$ denotes convolution with filter $\boldsymbol{h}$ and $\boldsymbol{A}^T$ is known as the backprojection operator.

In Fourier-space direct methods such as GridRec, both filtering and backprojection are performed in Fourier space, after which a fast Fourier transform (FFT) is used to convert the Fourier-space reconstruction to a real-space reconstruction. In addition, Fourier-space methods often use a windowing function to improve the accuracy of interpolation in Fourier space [13].

An important point to note is that, although the problem of inversion is well defined in the continuous setting, the discretized reconstruction formula (1.18) depends on the choice of discretization and interpolation. These choices are usually implementation-specific, which means that they differ across the various available open-source software implementations of direct algorithms, and contribute to quantitative differences between reconstructions from each implementation.

Direct methods usually result in poor reconstructions when noise in the data is high or data have been collected over a limited angular range. For such data, methods that solve the linear least-squares problem (1.7) iteratively are better. One popular iterative method is SIRT, which solves the linear least-squares problem with additional preconditioning and, optionally, non-negative constraints. For large data, a major limiting factor to the practical application of iterative methods is that the computation time required for reconstructing is much larger than the time required by direct methods.

A class of filter-optimization methods seek to augment direct methods with some of the advantages of iterative methods without compromising on the speed of computation. In such methods, the filter in direct algorithms ($\boldsymbol{h}$ in (1.18)) is learned from the data $\boldsymbol{y}$, following which they can be used on-the-fly with direct methods in place of standard hand-crafted filters. Filter learning using a minimum-residual approach has been performed for FBP [14] as well as the Feldkamp–Davis–Kress (FDK) algorithm [15], which generalizes FBP to cone-beam tomography setups.

A minimum-residual filter for data $\boldsymbol{y}$ can be computed by solving the following optimization problem:

$$\underset{\boldsymbol{h}}{\text{minimize}} \quad \|\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{A}^T \boldsymbol{C}_{\boldsymbol{y}}\,\boldsymbol{h}\|_2^2. \tag{1.19}$$

The minimum-residual filter $\boldsymbol{h}^* := \sum h_i^*\,\boldsymbol{b}_i$ is expressed as a linear combination of basis vectors $\boldsymbol{b}_i$. Sparse design of the filter is possible by choosing an appropriate basis such that only a few filter coefficients have to be learned. A binned basis with exponentially wider bins away from the centre of the detector array was first proposed for FBP [14] and later used for filter computation for FDK [15] without loss of reconstruction accuracy. Such an exponential binning basis translates to a basis of linear combinations of cosines in Fourier space.

In Chapter 2 we use such sparse-basis filters to tackle the problem of reproducibility in synchrotron tomography. Hardware and software vary across synchrotrons, and the results of experiments performed by users at different facilities are often not directly comparable with each other. We focus on the image reconstruction block in the synchrotron tomography pipeline, where differences between discretization and interpolation in various software packages play a role in enhancing differences between experimental results. We show that minimum-residual filters can improve the similarity between reconstructions (of synthetic and real data) obtained from several open-source implementations of direct algorithms, and contribute to a more reproducible reconstruction block in the synchrotron pipeline.

Sparsity is used to limit the number of filter coefficients of minimum-residual filters.

1.4.2 Grid-free, sparse reconstruction of nanocrystal defects

The goal of atomic-resolution electron tomography is to get a precise quantitative picture of a nanocrystal down to the atomic scale. To probe at such high resolutions, optimizing both the acquisition of projection images and the reconstruction of projection data is required.

One approach to atomic-resolution reconstruction is based on discrete tomography. In this approach, atoms are assumed to lie on a regular lattice and the measured projections are considered as atom counts along lattice lines. A key advantage of this approach is its ability to exploit the constraints induced by the discrete domain and range of the image. As a consequence, a small number of projection angles (typically less than 5) can already lead to an accurate reconstruction [16], [17]. A key drawback of the discrete lattice assumption is that in many interesting cases the atoms do not lie on a perfect lattice due to defects in the crystal structure or interfaces between different crystal lattices.

As an alternative, it has been demonstrated that a more conventional tomographic series consisting of hundreds of projections of a nanocrystal can be acquired in certain cases. An image of the nanocrystal is then reconstructed using sparsity-based reconstruction techniques on a continuous model of the tomography problem, typically solving the problem (1.16). This approach does not depend on the lattice structure and allows one to reconstruct defects and interfaces [18]. As a downside, the number of required projections is large and to accurately model the atom positions the reconstruction must be represented on a high-resolution pixel grid resulting in a large-scale computational problem. More importantly, increasing the resolution of the pixel grid in order to capture defects results in a much more ill-posed problem.

For the atomic-resolution reconstruction problem, a canonical discretization is provided not by an arbitrarily imposed pixel grid but by the spatial coordinates of atoms in the nanoparticle. Optimizing over a set of atom coordinates is nonconvex; a convex formulation proposed in the context of single-molecule localization microscopy [12] involves mapping the problem to the space of measures. In the space of measures, a set of atoms can be represented as a positive measure $\mu := \sum_{i=1}^{N_{\text{atoms}}} w_i \delta_{\boldsymbol{x}_i}$, with $\boldsymbol{x}_i$ being the spatial coordinates of atom i and $w_i \geq 0$ denoting weights that scale the intensity of atom i in projection data. Reconstructing the correct measure means solving

$$\underset{\mu}{\text{minimize}} \quad \|\Phi\, \mu - y\|_2^2, \tag{1.20}$$

where the forward model $\Phi\, \mu$ maps the measure to data y, such that

$$\Phi\, \mu := \mathcal{R}\Big(\sum_{i=1}^{N_{\text{atoms}}} w_i (G * \delta_{\boldsymbol{x}_i}) \Big),$$

where $\mathcal{R}$ is the continuous Radon transform and G denotes a known shape function. Sparsity can be included in the optimization problem in a number of ways: one way is by adding a term that minimizes the ℓ^1-norm of the weights $\{w_i\}$, another is by using a Frank–Wolfe-type algorithm [19] where the objective is minimized iteratively and only one atom is added to the support of the measure μ at each iteration.

In Chapter 3, we investigate grid-free algorithms to solve the reconstruction problem above. We demonstrate the advantages of using a grid-free approach to traditional grid-based reconstruction algorithms. We also show that including physical priors relevant to the problem – in this case, the potential energy of the atomic configuration – can help to resolve configurations with greater accuracy, especially in situations where the projection data are not enough to determine a unique atomic configuration.

Atomic configurations are modelled as sparse measures, whose support is the locations of atomic centres in continuous space.

1.4.3 Grid-free tilt-series alignment in cryoET

The goal of cryoET is to study the structure of biological macromolecules, such as proteins, in their native cellular context. Aspects of cryoET that distinguish it from other CT setups are as follows. Firstly, the geometry of the experimental system limits the extent to which the sample can be tilted. Moreover, the increase in apparent sample thickness with increasing tilt allows projection images to only be acquired for a limited angular range in cryoET, usually in $[-60^\circ, 60^\circ]$, resulting in a *missing wedge* of information that is not available during reconstruction [20]. Secondly, cryoET samples are dose-sensitive, which limits the total dose during acquisition and leads to very noisy projection images when a large number are acquired. Thirdly, the sample undergoes local and global movements during the acquisition procedure, making it difficult to reconstruct with a constant sample assumption.

Local deformation of the sample induced by the electron beam is a key resolution-limiting factor in cryoET. One way to align the tilt series is by using high-contrast gold beads as markers and modelling the deformation of markers using prior knowledge on sample deformation.

Extending the formalism of the previous section, the deforming marker configuration in cryoET can be mapped to a measure $\mu := \sum_{i=1}^{N_{\text{markers}}} w_i \delta_{\boldsymbol{x}_i}$ and the projection data at time t can be modelled using a forward model given by

$$\Phi_t \, \mu := \sum_{i=1}^{N_{\text{markers}}} w_i \, (G * \mathcal{R}) \, \delta_{\boldsymbol{x}_i + \boldsymbol{D}_t(\boldsymbol{P}, \boldsymbol{x}_i)}, \tag{1.21}$$

where $\boldsymbol{D}_t(\boldsymbol{P}, \boldsymbol{x}_i)$ denotes a deformation field with parameters $\boldsymbol{P}$. Tilt-series alignment then amounts to optimizing over the deformation parameters, marker locations and weights, and number of markers N_{markers}:

$$\underset{w_i, \boldsymbol{x}_i, \boldsymbol{D}_t, N_{\text{markers}}}{\text{minimize}} \quad \sum_{t=0}^{T} \left\| y_t - \sum_{i=1}^{N_{\text{markers}}} w_i \, (G * \mathcal{R}) \, \delta_{\boldsymbol{x}_i + \boldsymbol{D}_t(\boldsymbol{P}, \boldsymbol{x}_i)} \right\|_2^2 . \tag{1.22}$$

We tackle this problem in Chapter 4 of this thesis, and show that our grid-free formulation allows the recovery of deformation parameters in synthetic and real data accurately despite the absence of labelled marker data as in existing approaches.

Gold-bead markers are modelled as sparse measures that deform over time according to a parametrized deformation field.

1.4.4 Cell–cell interaction learning for vascular network formation

In the final application area studied in this thesis, we look at the problem of inferring cell–cell interactions that are necessary for vascular network formation. To do this we adopt a method called Sparse Identification of Nonlinear Dynamics (SINDy) that has been shown to recover dynamical equations from time-series data [21].

The SINDy approach is applicable to ordinary differential equations of the type:

$$\dot{\boldsymbol{x}} = \boldsymbol{g}(\boldsymbol{x}), \tag{1.23}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ denotes the system state at a certain time and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^n$ is a vector field that defines the dynamics of the system. Given measurements of $\boldsymbol{x}$ at a discrete set of time-points $\mathcal{T}$, and using computed values for $\dot{\boldsymbol{x}}$ at these time-points, the goal of SINDy is to recover the functional form of $\boldsymbol{g}$ from a library of functions. To do this, SINDy solves the following optimization problem:

$$\underset{\boldsymbol{\xi} \in \mathbb{R}^K}{\text{minimize}} \quad \|\dot{\boldsymbol{x}} - \Theta(\boldsymbol{x})\boldsymbol{\xi}\|_2^2 + \lambda \|\boldsymbol{\xi}\|_1 \,, \tag{1.24}$$

where $\Theta(\boldsymbol{x})$ denotes the library functions evaluated at the data points and $\boldsymbol{\xi}$ is the vector of coefficients that weights the library terms. Thus, SINDy optimizes for a sparse set of library terms that describes the measurements of a dynamical system.

SINDy has been used to infer the dynamics of simulated and real data for a variety of canonical systems exhibiting nonlinear dynamics [21]. Moreover, extensions of the SINDy approach have been used to investigate several problems of biological relevance. Two important examples are learning stochastic differential equations [22] and implicit ordinary differential equations describing biological networks [23].

In our case, the vascular network formation process is described by the dynamical equation (1.4) in particle-based simulations. In the overdamped regime, this translates to a form for $\boldsymbol{g}$ given by

$$g_i(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) := \sum_{j \in \mathcal{N}_i} \frac{\boldsymbol{x}_i - \boldsymbol{x}_j}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|} F_{ij} \,, \quad i = 1, \ldots, n_p \tag{1.25}$$

$$F_{ij} := \Phi(\boldsymbol{x}_i, \boldsymbol{x}_j, \gamma_i, \gamma_j), \tag{1.26}$$

where we parametrize the force between cell pairs as a function Φ of cell locations $\boldsymbol{x}_i$ and orientations γ_i. The learning problem then amounts to learning a form for these cell–cell interactions F_{ij} from a library of functions.

In Chapter 5 we provide details of how this can be done, and apply our learning approach to simulation studies of vascular network formation. Our method can be extended to recover similar interaction terms from experimental data, and enables the discovery of effective equations from observations of a few variables.

Sparsity is used to constrain the number of terms in the inferred pairwise interaction between cells.

1.5 Research questions

To close this introductory chapter, we present the four research questions that were investigated in this thesis. Each of these research questions is presented on a separate page and is dealt with in a separate chapter. Here we provide a brief abstract of our method and main contributions, along with a representative illustration.

Research question 1. *Can sparse-basis minimum-residual filters be used to improve reproducibility in the synchrotron CT pipeline?*

In Chapter 2, we propose a filter-learning approach that reduces the quantitative differences between reconstructions obtained from popular open-source implementations. These differences are a result of differing software conventions for discretization and interpolation. We show that optimizing the filter in real-space and Fourier-space direct algorithms reduces such differences, resulting in fewer differences also in post-processing results. We apply our method to real data acquired at the synchrotron to validate the usability of our approach.

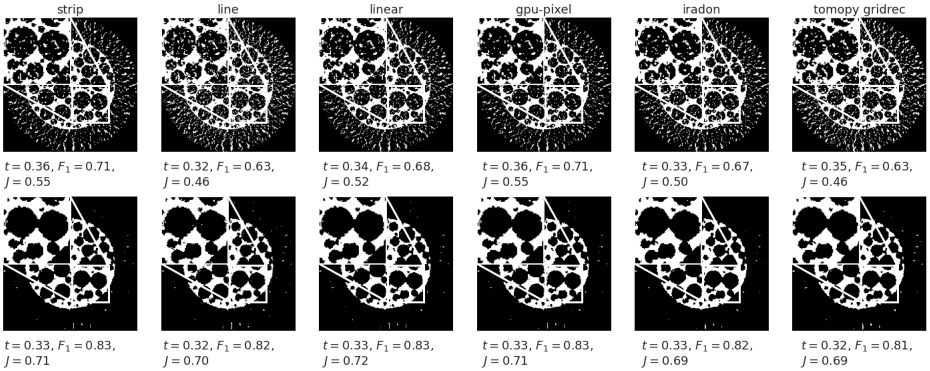

strip
line
linear
gpu-pixel
iradon
tomopy gridrec
$t = 0.36$, $F_1 = 0.71$, $J = 0.55$
$t = 0.32$, $F_1 = 0.63$, $J = 0.46$
$t = 0.34$, $F_1 = 0.68$, $J = 0.52$
$t = 0.36$, $F_1 = 0.71$, $J = 0.55$
$t = 0.33$, $F_1 = 0.67$, $J = 0.50$
$t = 0.35$, $F_1 = 0.63$, $J = 0.46$
$t = 0.33$, $F_1 = 0.83$, $J = 0.71$
$t = 0.32$, $F_1 = 0.82$, $J = 0.70$
$t = 0.33$, $F_1 = 0.83$, $J = 0.72$
$t = 0.33$, $F_1 = 0.83$, $J = 0.71$
$t = 0.33$, $F_1 = 0.82$, $J = 0.69$
$t = 0.32$, $F_1 = 0.81$, $J = 0.69$

Figure 1.4: Differences between post-processing results after thresholding with Otsu's method. The top row shows thresholded reconstructions obtained using different backprojector implementations and a standard Shepp-Logan filter; Otsu thresholds t, F_1 scores and Jaccard indices are given for each image. The bottom rows shows thresholded reconstructions obtained using our implementation-adapted filters. Both qualitatively and quantitatively these results are more similar to each other than those in the top row.

Research question 2. *Can grid-free sparse reconstruction approaches infer the locations of defects in nanocrystals from very few projections?*

In Chapter 3, we turn to the atomic-resolution ET problem and propose a grid-free sparse optimization approach to tackle it. We also show how adding a physical potential energy term to the optimization objective helps to resolve atomic configurations from only two or three projections. We compare the performance of our method with that of existing grid-based methods such as SIRT and FISTA, as well as with that of nonconvex techniques like simulated annealing.

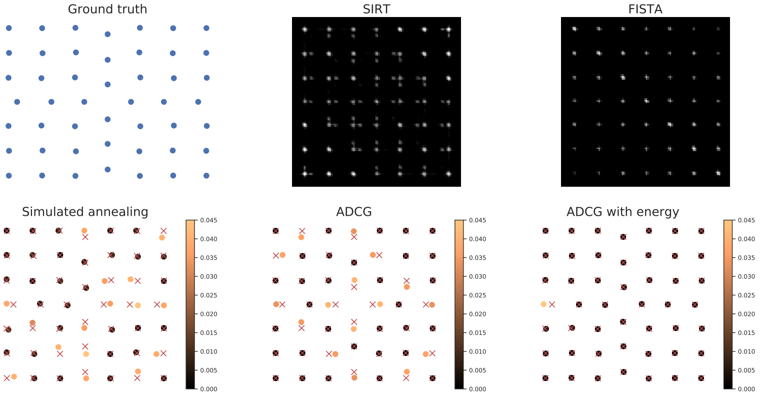
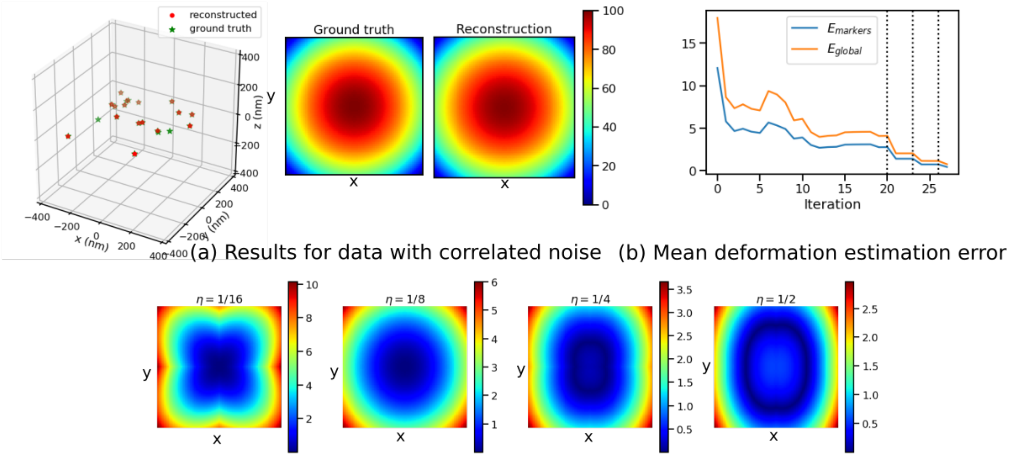


Figure 1.5: Reconstructions of a vacancy defect from three projections. For the simulated annealing, ADCG and ADCG with energy reconstructions, atoms are coloured according to their Euclidean distance from the ground truth. Ground truth positions are marked with red crosses.

Research question 3. *Can we extend grid-free sparse optimization methods to infer deformation parameters for cryoET alignment?*

In Chapter 4, we extend and adapt a grid-free algorithm to infer both locations and deformation parameters of gold markers in cryoET. We use globally supported parametrized deformation fields based on previous experimental studies to model beam-induced sample motion. The parameters of this model and marker locations are simultaneously inferred from our method, without the need for labelled marker data in each projection. We apply our method to TEM simulations as well as real data of gold beads on ice, and show that our method can estimate deformation fields in a host of noise and model mismatch settings.



(a) Results for data with correlated noise (b) Mean deformation estimation error

(c) Deformation estimation error at different resolutions

Figure 1.6: Inference of marker locations and deformation parameters from simulated TEM data with correlated noise. (a) Reconstructed and ground truth marker locations (left), and reconstructed and ground truth deformation fields in the direction of the electron beam (right). (b) Deformation estimation error as a function of iterations. (c) Deformation estimation errors in the beam direction

Research question 4. *Can sparse equation learning recover cell–cell interactions from simulated time-series data of vascular network formation?*

In Chapter 5, we adapt a sparse equation-learning approach to infer which pairwise interaction terms contribute to vascular network formation. We run particle-based simulations of network formation to generate cell trajectories over time. We formulate the time evolution of the system to be given by an overdamped Langevin equation with force terms that correspond to the pairwise interactions between cells. These force terms are then inferred from the cell trajectory data from a library of plausible forces.

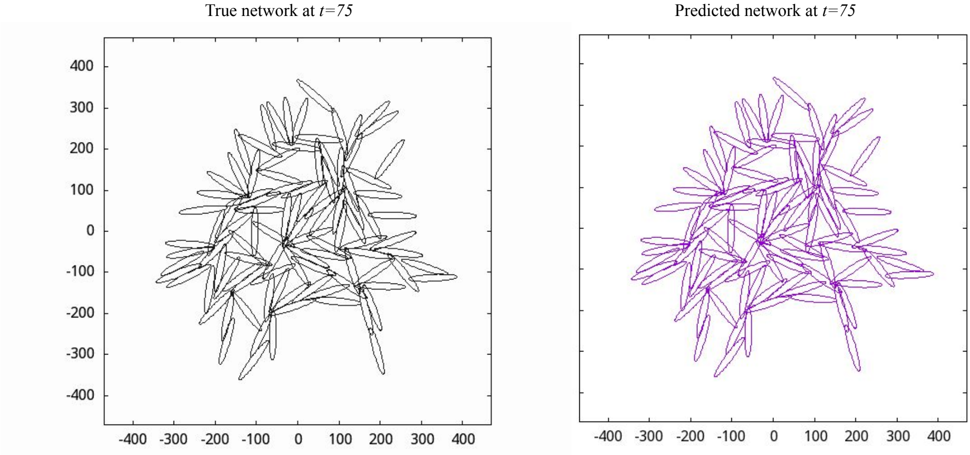


Figure 1.7: True and inferred vascular networks using 100 elongated cells.