Martin Zach

Generative Regularizers in Computed Tomography

MASTER'S THESIS

to achieve the university degree of
Diplom-Ingenieur

Master's degree programme
Biomedical Engineering

submitted to

Graz University of Technology

Supervisor

Prof. Dr. Thomas Pock
Institute of Computer Graphics and Vision

Dr. Erich Kobler
Institute of Computer Graphics and Vision

Graz, Austria, June 2021

Abstract

In today's medical landscape, imaging systems are of exceptional importance. Magnetic Resonance Imaging and Computed Tomography (CT) are heavily used and have improved diagnostic capabilities in many fields. However, contrast in *CT* relies on depositing ionizing radiation in the body, the dose of which should be as low as possible. Image quality per dose has improved drastically in the past decades, with advances in instrumentation and reconstruction algorithms.

In this work, we continue this trend by introducing a novel regularization scheme, where a parametrized regularizer is learned on data using maximum likelihood. Our energy-based formulation allows for much improved interpretability when compared to traditional feed-forward approaches. We can draw samples from our prior as well as the posterior of any given reconstruction problem, such that domain experts can judge the regularizer. We apply the regularizer to typical reconstruction tasks such as limited-angle and few-view *CT* reconstruction. Our model outperforms traditional reconstruction algorithms by a large margin.

Keywords. Deep Learning, Variational Methods, Data-driven Regularizers, Maximum Likelihood, Computed Tomography, Inverse Problems

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used.

The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Date	Signature

Acknowledgments

First and foremost, I want to thank my supervisors Thomas Pock and Erich Kobler. During the journey of writing this thesis, we had many fruitful discussions and their exceptional understanding of the matter helped me a great deal. Further, I want to thank all other colleagues of mine at the Institute of Computer Graphics and Vision for making my time there not only interesting, but also fun. Finally, I want to use this opportunity to thank my family, and especially my parents for enabling me to have such a carefree time of studying.

Contents

List of Figures

List of Tables

1 Introduction

> If you wish to make an apple pie from scratch, you must first invent the universe.
>
> Carl Sagan

In the last decade, deep learning [35] has taken over the field of computer vision, where learning-based approaches have improved image quality in restoration tasks such as denoising [103] or deblurring [70], and accuracy in tasks such as classification [57] or semantic segmentation [21]. In medical imaging, deep learning has traditionally been used as a tool to aid interpretation of reconstructed images, for instance through automatic segmentation [49] or classification [63]. However, to increase the visual quality of medical images, learning-based approaches may also be used at earlier stages such as during data acquisition or image reconstruction.

Computed Tomography (CT) images have historically been reconstructed using the fast Filtered Back-Projection (FBP) [12]. However, the analytical *FBP* has been superseded by iterative algebraic reconstruction algorithms at around the start of the new millenium [86, 98]. This drift has been driven largely by an increase in computational power and the need for more robust reconstruction algorithms in the light of reducing the administered ionizing radiation dose. In general, dose reduction is one of the major concerns in the *CT* community [17, 102]. It has been estimated that up to 50 % of ionizing radiation exposure for medical use can be attributed to *CT* examinations [71]. Thus, it is important to find reconstruction algorithms that are able to reconstruct a clinically valuable image from low-dose measurements, which may only contain a subset of the full-dose scan, or exhibit low Signal-to-Noise Ratio (SNR).

To make the iterative algebraic reconstruction algorithms more robust, prior knowledge about the solution may be incorporated in the reconstruction problem. Traditional, hand-crafted priors, such as the Total Variation (TV) prior [85] and extensions such as the Total Generalized Variation (TGV) [9], typically encode local gradient information of the reconstruction. While these hand-crafted priors have been used extensively and successfully in image restoration [16,

33, 85] and reconstruction tasks [23, 62, 105], they lack expressiveness compared to state-of-the-art learning-based approaches [56]. In a similar vein, traditional learning-based approaches model local image information and therefore are not suited to fully remove the global streaking and smearing artifacts that arise can arise low-dose and limited-angle *CT* [4]. A popular approach to combat this issue has emerged recently and is based upon learning the stages of an iteratively unrolled gradient descent individually [40, 41]. However, although this approach does consider physical principles, it lacks interpretability due to its feed-forward nature. In contrast, we propose a novel learned prior utilizing a global receptive field to remove large-scale coherent artifacts, while staying consistent with the acquired data.

1.1 Contributions and Outline

In this thesis, we introduce a novel regularization scheme, where a regularizer with a global receptive field is trained generatively on *CT* images. Our formulation allows to cast the regularizer in a probabilistic framework, which drastically improves interpretability compared to other deep learning-based approaches. As an example, we can visualize the prior distribution of our regularizer by showing its modes or drawing samples from it. Further, for any given reconstruction problem, in addition to computing the Maximum-A-Posteriori (MAP) point estimate, the posterior distribution can be sampled. Therefore, the expected value as well as the variance over the posterior can be visualized, which is valuable as it relates to uncertainty quantification.

We apply a trained model to a multitude of reconstruction tasks, and compare our approach quantitatively and qualitatively with more traditional reconstruction algorithms. In addition, we perform experiments which leverage the possibility of probabilistic interpretation of our approach, such as prior and posterior sampling. Finally, we challenge our proposed novel approach by applying our regularizer to reconstruction tasks of different resolutions and out-of-distribution data. To summarize, we enumerate the contributions in this thesis as follows:

- The design of an architecture that is suitable for usage as a generative prior.
- Data-independent analysis of the learned regularizer by means of exploring modes and drawing samples, for visualizing and understanding the learned regularizer.
- Application of the learned regularizer to limited-angle and few-view *CT* reconstruction, achieving satisfactory results. In addition, we analyze the posterior distribution of a few-view reconstruction problem.
- Pointing out the limitations of our approach by challenging the learned regularizer on out-of-distribution data.

This thesis is organized as follows: In Chapter 2, we will introduce the general principle of tomography and review the physical principles in medical *CT*. Along the way, we will develop the signal model for *CT* that is used throughout this thesis. We end this chapter with a brief overview of medical *CT* instrumentation. In Chapter 3, we develop a mathematical formulation

for the forward problem and discuss approaches for solving the inverse problem. We also review the typical artifacts that arise during reconstruction. In Chapter 4, we discuss different possibilities for increasing the visual quality of *CT* reconstructions. This ranges from pre- and post-processing techniques over domain transform learning to variational reconstruction, which is the approach we follow. We discuss the specifics of our proposed model and training procedure and show the corresponding numerical experiments in Chapter 5. Finally, the thesis is concluded in Chapter 6.

2

Principles of X-Ray Computed Tomography (CT)

I did not think, I investigated.

Wilhelm Röntgen, *McClure's Magazine VI No. 5, 1896*

Contents

Modern medicine relies greatly on imaging techniques, where different physical processes are exploited to give insights into the human body. For instance, Magnetic Resonance Imaging (MRI) measures proton density with the help of strong magnetic fields, field gradients and radio pulses, by which an excellent soft-tissue contrast can be achieved. The importance of *MRI* in medicine is emphasized by the 2003 Nobel Prize in Physiology or Medicine, which was awarded to Peter Mansfield and Paul Lauterbur "for their discoveries concerning magnetic resonance imaging"[1].

X-Ray *CT* is of similar importance in modern medicine, where by help of ionizing radiation cross-sections of the body are imaged with great hard-tissue contrast. In 1979, Allan Cormack and Godfrey Hounsfield were rewarded the Nobel Prize in Physiology or Medicine "for the development of computer assisted tomography"[2]. Practically, the advantages of *CT* over *MRI* are the reduced costs in both acquisition and operation, as well as faster image acquisition. The main disadvantage is the usage of ionizing radiation, whose energy fundamentally must be at least partly deposited in the body to create contrast.

In the following sections, we will outline the principles of X-Ray *CT*. We will define the tomography problem, go over physical principles and instrumentation in *CT*, and will build the

[1] see `https://www.nobelprize.org/prizes/medicine/2003/summary/`, accessed 2021-04-19

[2] see `https://www.nobelprize.org/prizes/medicine/1979/summary/`, accessed 2021-04-19

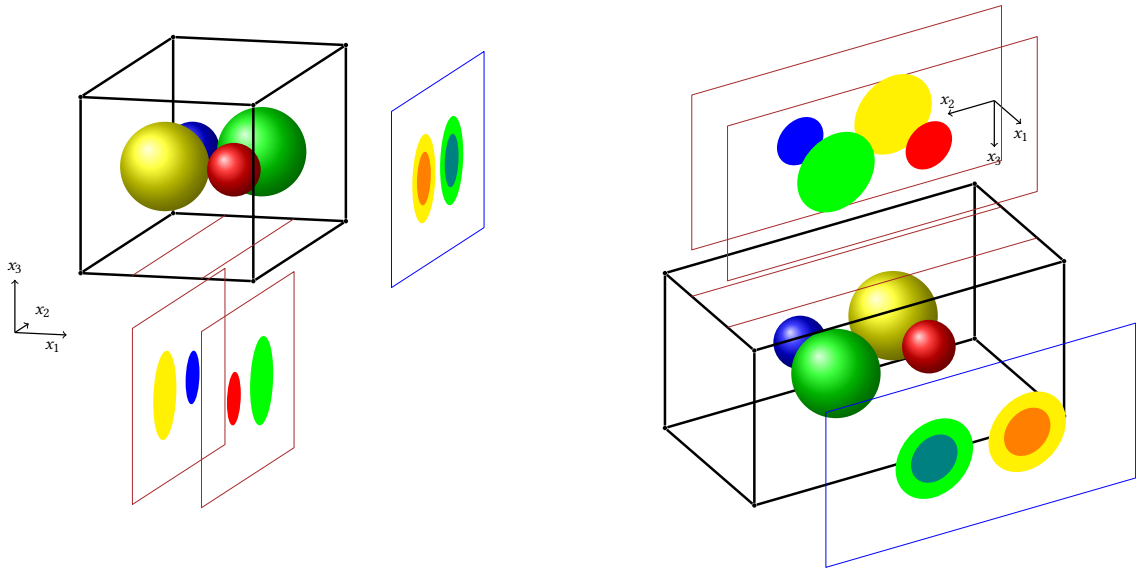


Figure 2.1: Tomography versus projectional imaging: In tomography, the goal is to acquire the cross sections shown in red, whereas in projectional imaging, a specific direction is integrated over.

mathematical foundation needed in the next chapters along the way. Since our focus is on reconstruction, we refer the reader to [12, 76] for a more in-depth review of the physics of *CT*.

2.1 Tomography

In the most general sense, *tomography* describes the "imaging of cross-sections". The word is derived from the Greek words τόμος (tomos, "slice" or "section") and γράφω (graphō, "to write" or "to describe"). Specifically, given some volume, we are interested in visualizing distinct slices with minimal interference of the rest of the volume. This is distinctly different from projectional methods, where the resulting two-dimensional projection displays information of the volume integrated over a specific direction. We show an example illustrating the difference between tomography and projectional imaging in Fig. 2.1.

Besides medical applications, where the body of interest is (a specific part of) the human body, tomography is widely used in the geosciences, where the body of interest is (a specific part of) the earth. As an example, in seismic tomography, seismographs across the earth's surface register motion of the ground induced by earthquakes. With this information, certain characteristic of the rock can be reconstructed [75]. In muon tomography, muons from cosmic radiation are used to image large-scale structures. This has been used to image the reactor after the 2011 Fukushima Daiichi nuclear disaster, to assess the situation of the remains of the reactor cores [7].

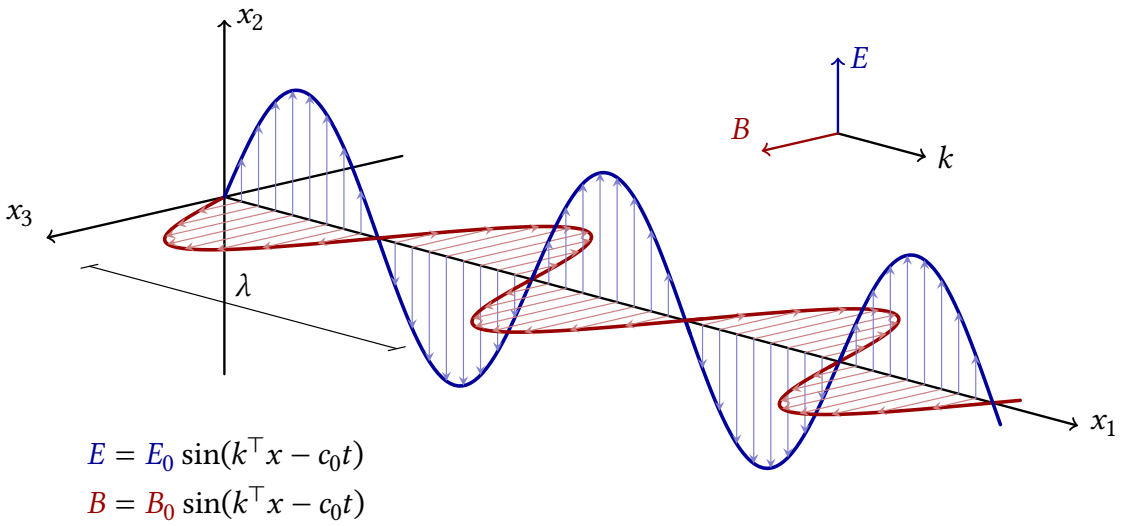


Figure 2.2: The propagation of an electromagnetic wave: The electric field E is perpendicular to the magnetic field B. The direction of propagation is referred to as the wave number k.

2.2 Physical Principles

The underlying physical principle of X-Ray *CT* and X-Ray projection radiography is attenuation of electromagnetic radiation by any given medium. The term "X-Ray" itself describes a specific interval in the electromagnetic spectrum, which is of high enough energy to be classified as *ionizing* radiation. Ionizing radiation, as opposed to non-ionizing radiation, is capable of ejecting electrons from atoms and thereby creating ions. It is the interaction between ionizing radiation, most often produced by the X-Ray tube, and the atoms of the patient's body that ultimately yield the contrast in the X-Ray image.

2.2.1 Electromagnetic Radiation

Classically, electromagnetic radiation describes waves of the electromagnetic field propagating through space. An electromagnetic wave consists of an electric and a magnetic component, which are perpendicular to each other. The relationship between the electric and magnetic component is described by the famous Maxwell's equations. We schematically show the propagation of a linearly polarized electromagnetic wave with wavelength λ in Fig. 2.2. The wavelength of a electromagnetic wave is related to its frequency by

$$\nu = \frac{c_0}{\lambda} \tag{2.1}$$

with the speed of light $c_0 = 3 \times 10^8\,\mathrm{m\,s^{-1}}$.

The particles associated with electromagnetic radiation are referred to as *photons*. The energy

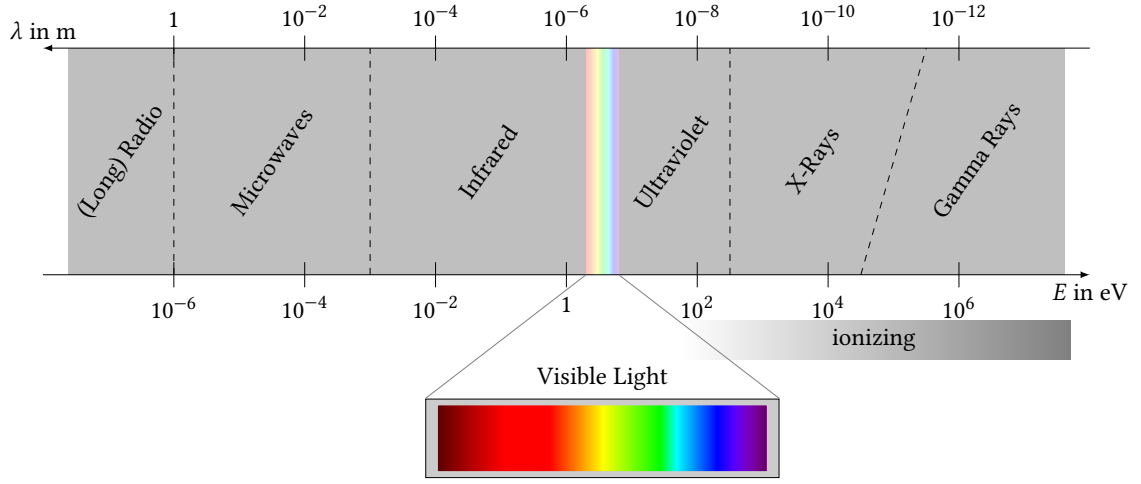


Figure 2.3: The electromagnetic spectrum with some frequently distinguished energy ranges. The slanted line should emphasize that X-Rays and Gamma rays are distinguished by their point of origin rather than their energy, and that there some overlap in energy.

of a photon associated with a wave of frequency ν is

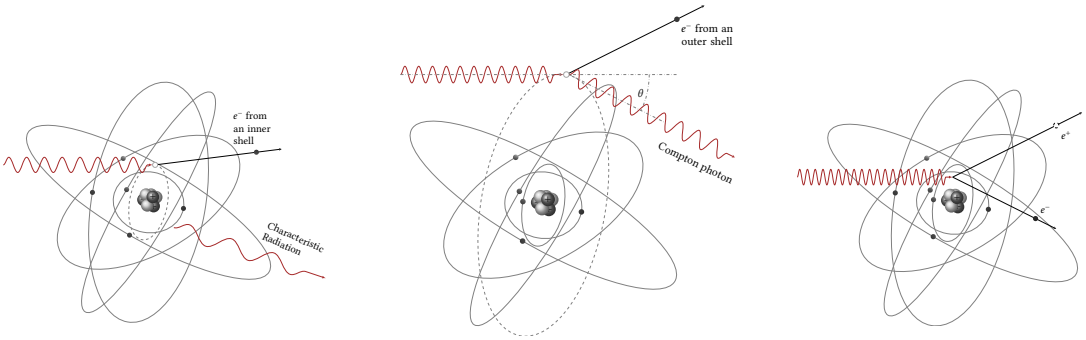
$$E = h\nu, \tag{2.2}$$

where $h = 6.626\,070\,15 \times 10^{-34}$ J s is Planck's constant. Electromagnetic waves exist on an energy spectrum, where different classes are defined. Radio waves, visible light and X-Rays are examples of electromagnetic radiation, with radio waves having the lowest energy of the three. For medical radiography applications, radiation in the range of 25 keV to 500 keV is used.

We show the electromagnetic spectrum with some important classes in Fig. 2.3. We note that, although X-Rays and Gamma rays are usually distinct classes in the electromagnetic spectrum, they are not distinguished by energy but by point of origin. Specifically, X-Rays are defined as radiation originating from the electron cloud, while gamma rays originate from the nucleus of an atom.

2.2.2 Interactions of Electromagnetic Radiation with Matter

Ionizing electromagnetic radiation interacts with matter primarily by 1. the photoelectric effect, 2. Compton scattering, and 3. pair production. Typically, pair production is only considered relevant for high energy photons with $E > 1.022\,\text{MeV}$. As previously mentioned, in the medical imaging domain the highest energies are approximately 500 keV and as such pair production can largely be neglected. For both the photoelectric effect and Compton scattering, the interaction of the X-Ray with the atoms happens in the electron shell. The defining difference between the photoelectric effect and Compton scattering is that in the latter (the energy of) the photon is not fully absorbed by the atom. In medical imaging, the contrast is mostly due to the photoelectric effect, whereas Compton scattering limits the resolution of the images. We show the three



(a) Photoelectric Effect (b) Compton Effect (c) Pair Production

Figure 2.4: Illustration of the three principles of interaction between electromagnetic radiation and matter. The wavelength of the incident ray indicates the relative energies that these effects are most likely to occur at. Pair Production only happens for $E > 1.022\,\text{MeV}$, and as such can largely be ignored for medical applications.

interaction mechanisms schematically in Fig. 2.4.

2.2.2.1 Photoelectric Effect

In the photoelectric effect, through interaction of a photon with the coulomb field of the nucleus, an electron from an inner shell, most often the K-shell, is ejected from the atom. The incident photon of energy $E_p = h\nu_p$ is fully absorbed by the atom, and the electron is ejected with energy

$$E_{e^-} = E_p - E_B, \tag{2.3}$$

where E_B is the binding energy of the electron. The hole of the ejected (usually K-shell) electron is then filled by an electron in an outer shell. The energy difference between the shells is converted into electromagnetic radiation (X-Rays), which is characteristic to the atom, as the structure of the atom dictates the energy delta between the shells.

The resulting characteristic X-Rays may sometimes eject another outer-shell electron, called the *Auger electron*, and consequently lead to readjustment of the remaining electrons. The ejected electrons (photoelectrons and Auger electrons) can then be treated as typical particulate radiation, and interact with the matter around them. In fact, these particles contribute largely to the biological effects of ionizing radiation. We show an illustration of the photoelectric effect (without considering Auger electrons) in Fig. 2.4a.

2.2.2.2 Compton Scattering

In contrast to the photoelectric effect, in Compton scattering the energy $E_p = h\nu_p$ of the incident photon is not fully absorbed by the atom. Instead, it loses some energy in the process of ejecting an outer-shell electron (the *Compton electron*), and is deflected by the *Compton angle* θ. These concepts are illustrated in Fig. 2.4b. Depending on the Compton angle θ, the energy E_c of the

Compton photon is given by

$$E_c = \frac{E_p}{1 + (1 - \cos\theta)\frac{E_p}{{m_e c_0}^2}} \tag{2.4}$$

where ${m_e c_0}^2 = 511\,\text{keV}$ is the rest energy of an electron. We see that the Compton photon has the smallest energy when $\theta = \pi\,\text{rad}$, i.e. when the photon is reflected back to the incidence direction. The ejected electron is again free to interact with the surrounding matter.

2.2.2.3 Pair Production

Although not interesting for medical applications, we briefly discuss pair production here for completeness. Pair production is the primary principle of interaction of high-energy photons with matter. Specifically, incident photons with $E_p > 1.022\,\text{MeV}$, or $\lambda < 1.2132\,\text{pm}$, may "decay" into an electron-positron pair when near a nucleus. The requirements are very specific: The energy for such interactions is bounded from below with 1.022 MeV, as this is the rest energy $2{m_e c_0}^2$ of an electron-positron pair. Further, the decay event must happen near a nucleus, as otherwise conservation of momentum would necessarily need to be violated. As a result, the nucleus typically experiences some "recoil" during such events. The electron and positron are then free to interact with their neighborhood. The most likely fate of the positron is an electron-positron annihilation event, which can be thought of as the inverse of pair production.

2.2.3 Attenuation of Electromagnetic Radiation in Matter

In the previous sections, we described the primary principles by which electromagnetic radiation can interact with matter. Here, we want to consider these effects on a macroscopic level. We do this by introducing probabilistic concepts modeling the effects on a macroscopic level. With simplified examples, we finally arrive at a signal model in dependence of the quantity of interest, which is the attenuation coefficient of the imaged body.

2.2.3.1 Probabilities of Interaction

Let us first informally consider the probabilities of the different interactions of electromagnetic radiation with matter. The photoelectric effect requires interaction with the Coulomb field of the nucleus of an atom. Consequently, the probability of a photoelectric event is related to the atomic number Z of an atom. In a heterogeneous material composed of different elements, we define the effective atomic number Z_{eff}, which summarizes the compound. The probability of a photoelectric event is $\propto Z_{\text{eff}}^4$ for the compounds typically found in human tissue. Further affecting the probability of photoelectric events is the energy of the incident ray, such that in summary $\mathbb{P}(\text{PE event}) \propto \frac{Z_{\text{eff}}^4}{(h\nu)^3}$. The binding energy of the inner shell electrons also affects the probability of photoelectric events, which is exploited in contrast agents.

Compton scattering on the other hand describes an interaction with outer, loosely bound electrons. The probability of Compton events is mainly dependent on the electron density in the

Table 2.1: Relative impact of Compton Scattering versus the photoelectric effect for different photon energies in water. Adapted from [46].

Photon Energy in keV	Compton Interactions in %	Energy Transfer by CS in %
20	26.4	1.3
40	77.9	19.3
60	93.0	55.0
80	97.0	78.8
100	98.4	89.6
150	99.5	97.4
400	99.9	99.9

material, i.e. $\mathbb{P}(\text{CS event}) \propto \frac{N_A Z}{W_M}$. Here, $N_A = 6.022\,140\,76 \times 10^{23}\,\text{mol}^{-1}$ is Avogadro's constant, and W_M is the molecular weight in $\text{g}\,\text{mol}^{-1}$. In human tissue, the electron density does not vary hugely, with it being around $3.1 \times 10^{26}\,\text{kg}^{-1}$ [99] for air, water, muscle, fat and bone. Therefore, the probability of Compton events is largely independent of the atomic number. Also, according to the Klein-Nishina formula [55], the probability for Compton events in the energy range for medical applications is close to constant.

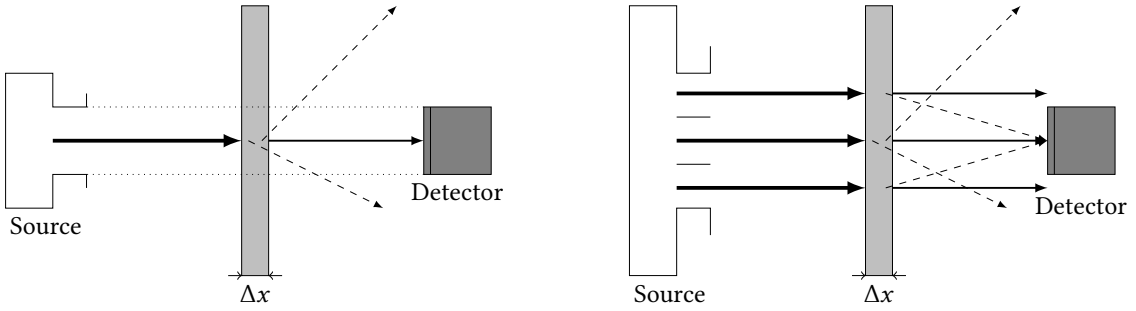
We show the relative frequency of occurrence between Compton and photoelectric events in Table 2.1 (adapted from [46]). It shows that in the energy range that is interesting for medical applications, except for the lower end, Compton scattering is the dominating modus of interaction in terms of number of occurrences. The share Compton scattering increases with increasing photon energy, where at $E_p = 60\,\text{keV}$ Compton scattering accounts for 93 % of all interactions. However, since in a Compton scattering event the energy is only partly deposited in the atom, it accounts for 55 % of the deposited energy at 60 keV. For energies $> 60\,\text{keV}$, Compton scattering also quickly becomes the dominating mode of energy transfer.

2.2.3.2 Monochromatic Narrow Beam and Homogeneous Thin Slab

To build up a signal model, we will consider the following setup: From some source, N photons with the same energy are shot perpendicularly onto a thin, homogeneous slab of thickness Δx. Behind the slab is a "perfect" detector with the same footprint as the incident beam. Clearly, without the slab we would expect the detector to count N photons. The slab however will absorb some photons by the photoelectric effect, as well as scatter photons by Compton scattering events such that they are no longer counted by the detector. In general, the detector will count $N' < N$ photons, i.e. the slab has *attenuated* the photon beam. We show the narrow beam geometry along with the distinctly different broad beam geometry in Fig. 2.5.

We denote with $\Delta N = N - N'$ the difference between the number of incident and detected photons. Assuming Δx is small, we expect ΔN to be proportional to both Δx and N, i.e.

$$\Delta N = \mu N \Delta x, \tag{2.5}$$



(a) Narrow Beam (b) Broad Beam

Figure 2.5: The narrow beam geometry a, where the footprint of the detector is as large as the incident beam, is a simple model for studying attenuation. If the incident beam is broader than the detector b, it may pick up scattered photons.

where μ is the (material dependent) proportionality constant, known as the *linear attenuation coefficient*. For the sake of simplicity, we treat N as a continuous quantity to quickly arrive at the well known differential equation

$$\frac{\mathrm{d}N}{N} = -\mu \, \mathrm{d}x, \tag{2.6}$$

such that

$$N' = N \exp(-\mu \Delta x), \tag{2.7}$$

where N is the number of photons at $\Delta x = 0$. In the monochromatic case, this immediately translates to

$$I' = I \exp(-\mu \Delta x), \tag{2.8}$$

with the intensity of the incident beam I.

2.2.3.3 Heterogeneous Slab

If the linear attenuation coefficient is not constant in the slab but varies with x, we modify Eq. (2.6) to

$$\frac{\mathrm{d}N}{N} = -\mu(x) \, \mathrm{d}x, \tag{2.9}$$

such that

$$N'(x) = N \exp\left(-\int_0^x \mu(\chi) \, \mathrm{d}\chi\right), \tag{2.10}$$

or, for the intensity

$$I'(x) = I \exp\left(-\int_0^x \mu(\chi) \, \mathrm{d}\chi\right). \tag{2.11}$$

This is known as the integral form of the fundamental X-Ray attenuation law and is the basis of the physical signal models for projected radiography as well as computed tomography.

2.2.3.4 Polychromatic Photons

As will be discussed later, X-Ray sources in medical imaging are polychromatic. We can extend the monochromatic narrow beam experiment to the polychromatic case by considering each energy independently, as the principles apply in general. The model then needs to account for the fact that the linear attenuation coefficient of human tissue does depend on the energy of the incident electromagnetic rays. Informally, let $S(E)$ be the intensity of the incident ray at energy E (i.e. the spectrum), and let $S'(x, E)$ be the spectrum at position x. We can modify the integral fundamental X-Ray attenuation law to account for the polychromatic photons as

$$S'(x, E) = S(E) \exp\left(\int_0^x \mu(\chi, E)\, \mathrm{d}\chi\right), \tag{2.12}$$

and consequently compute the intensity as

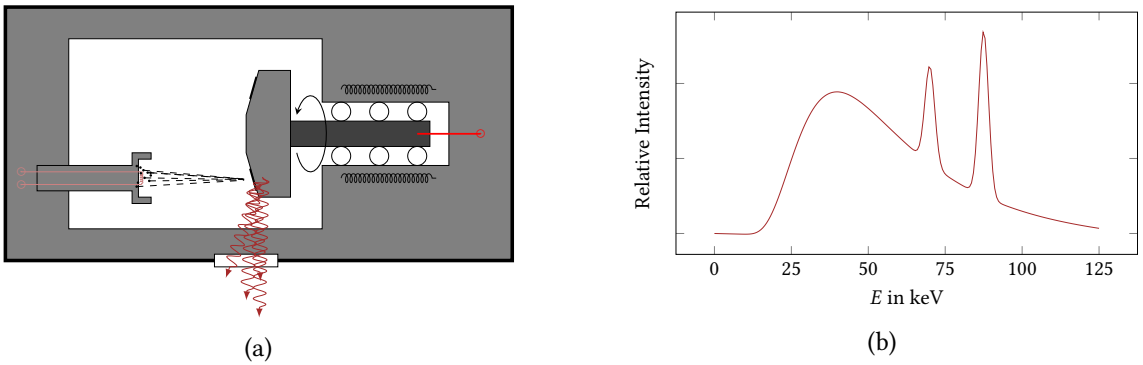
$$I'(x) = \int_0^{E_{\max}} S(\epsilon)\epsilon \exp\left(-\int_0^x \mu(\chi, \epsilon)\, \mathrm{d}\chi\right) \mathrm{d}\epsilon. \tag{2.13}$$

At this point we want to note that our goal is to reconstruct the position and energy-dependent linear attenuation coefficient $\mu(x, E)$. We can measure I' (at the detector), and typically $S(E)$ is known by the setup or by calibration measurements. However, solving Eq. (2.13) for $\mu(x, E)$ is in general intractable. To alleviate this, most often the concept of the *effective energy* is introduced, which is defined as the "the energy that, in a given material, will produce the same measured intensity from a mono-energetic source as is measured using the actual poly-energetic source".

2.2.3.5 Broad Beam

Let us now analyze the broad beam geometry seen in Fig. 2.5b. Clearly, without the slab we again expect N photons to reach the detector, as the other photon bursts will miss the detector completely. With the slab in place, we have the additional possibility that photons from other beams will scatter into the detector. As such, generally more photons reach the detector and the fundamental X-Ray attenuation law no longer holds. In addition, since the scattered photons have lost some energy in the scattering events, the detected photons are not mono-energetic, even if the incident beam was. Since the detected energies are shifted to the lower end, this is often referred to as *beam softening*.

Concretely, in terms of imaging systems, this is remedied by using collimator systems, such that the influence of photons from diverging directions is minimized. As such, for the mathematical analysis of the imaging signals, the narrow beam model is further assumed. However, we note that this simplification can not be made for determining the dose deposited in the patient or for calculating appropriate shielding of external personnel.



(a) (b)

Figure 2.6: The X-Ray tube in a shows the cathode assembly with the tungsten filament, where the electrons • are expelled by thermionic emission, and accelerated towards the anode. At the anode, bremsstrahlung and characteristic radiation produce a spectrum similar to b, which can exit the tube through a (filtered) window.

2.3 Instrumentation

In the previous section, we outlined the low-level physical interactions between electromagnetic radiation and matter and built a macroscopic signal model. In this section, we will discuss how *CT* systems are built in practice. We will again put particular focus on the abstract signal acquisition process rather than the physics of the X-Ray tube, filters, or detectors, although we discuss the main principles briefly.

2.3.1 X-Ray Tubes and Filters

In X-Ray *CT*, the X-Rays are generated by means of an X-Ray tube. On a basic level, the X-Rays used for diagnosis are the *bremsstrahlung* of previously accelerated electrons. There are two stages to this: First, the electrons are freed from some filament (usually Tungsten) in the cathode assembly by means of *thermionic emission*. Then, they are accelerated towards the anode (usually a Tungsten-Rhenium alloy on a Molybdenum core) by the *tube voltage*, which is in the range of 30 kV to 150 kV at its peak. At the anode, the high-energy electrons interact with the bulk matter, whereby bremsstrahlung and characteristic radiation is produced.

Note that the electromagnetic radiation can almost be considered a by-product in this process, as $\approx$ 99 % of the energy of the electrons is turned into heat rather than electromagnetic radiation. For this reason, proper cooling of the anode is essential. To help cooling, in most X-Ray tubes used for medical applications today, the anode is implemented as a rotating disk such that the energy of the electrons is dissipated over a larger area. Even with a rotating anode and active cooling, the focal spot can reach temperatures of up to 2000 °C, hence proper construction is vital. We show a sketch of an X-Ray tube along with a typical spectrum in Fig. 2.6.

The maximum energy of the bremsstrahlung is determined by the tube voltage, as it bounds the energy of the electrons from above. However, as seen in Fig. 2.6b, there exists a spectrum of lower energy photons that radiate away from the anode. Recall that, fundamentally, the contrast

in X-Ray *CT* arises from differential attenuation in the body. In other words, rays that are fully absorbed in the body or traverse the body without any interaction do not have any diagnostic value, but only contribute to the radiation dose of the patient. Therefore, the spectrum is usually filtered, such that low energy photons do not reach the patient. Note that the anode itself, the tube housing, and the cooling oil already largely filter low energy photons. In addition, a metal sheet can be placed outside the tube. Aluminium is most commonly used for this purpose, and other materials are described by "aluminium equivalent", i.e. the thickness of an aluminium sheet that would yield the same effect.

2.3.2 X-Ray Detectors

In most modern medical X-Ray *CT* scanners, solid state X-Ray detector lines or arrays are used, whereby high energy photons are converted into visible light by scintillation. The visible light is then converted into electric current by means of a photo-diode, and immediately amplified with a photomultiplier tube. Xenon gas detectors are also used, where thin tubes filled with Xenon generate a current between the cathode and the anode upon ionization of the gas by X-Rays. Although generally less efficient than their solid-state counterparts, they have the advantage of being highly directional, which is required in *CT* scanners of the third generation.

2.3.3 Scanner Generations

Recall that, in medical applications, our goal is to reconstruct the position and photon energy dependent linear attenuation coefficient $\mu(x, E)$ in the human body. We do this by measuring the attenuation along "all" lines between the source and the detector in a given cross section. In this section we will give an overview of different scanner generations, which achieve the above with different principles. Up to 7 generations of *CT* scanners are sometimes distinguished today. We show a summary of the scanner generations in Table 2.2, and illustrate the first four generations in Fig. 2.7.

First Generation First-generation scanners have the simplest geometry, which conceptually corresponds nicely to the mathematical theory of reconstruction discussed later. Specifically, they consist of a single collimated (i.e. "pencil beam") source along with a detector, that move linearly to acquire the attenuation along a line of a given angle. After one such linear scan, the source-detector assembly is rotated incrementally, and continues to acquire the attenuation as described before, as illustrated in Fig. 2.7a. Since the source and the detector move along a linear path (as opposed to, e.g. a fixed detector array), an arbitrary number of rays can be acquired. Analogously, the angle increment may be chosen arbitrarily small, such that one can freely choose the number of projections. This geometry further has the obvious benefit of conforming fully to the narrow beam geometry, i.e. the detected intensity only depends on the tissue along the path of the ray. The obvious disadvantage of this simple approach is the slow speed, which is why they are not used in clinical practice today.

Table 2.2: Overview of different scanner generations.

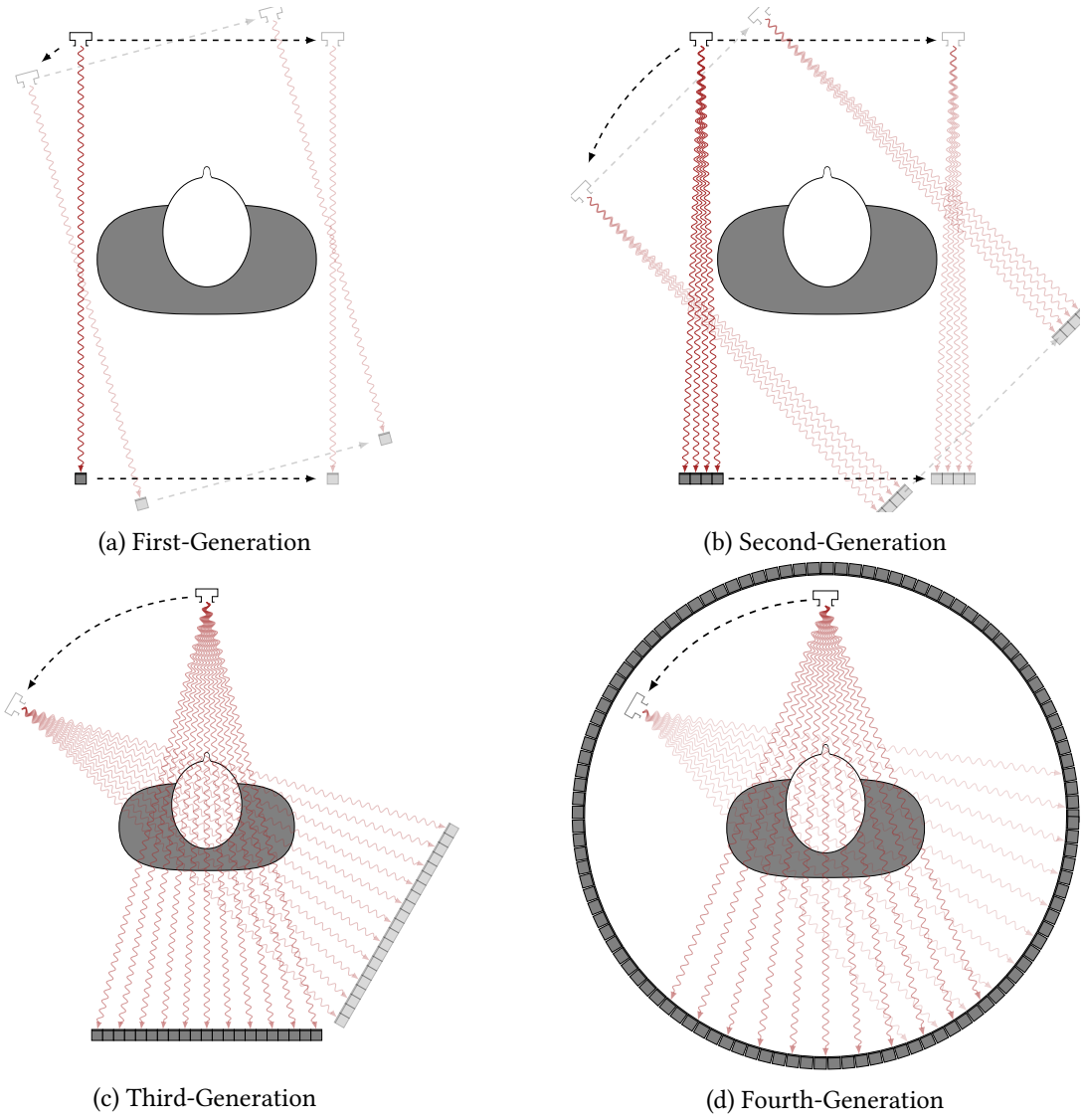
Scanner	Source	Detector	Coll.	Movt.	Advantages	Disadvantages
1G			✓*	$\{\rightarrow, \circlearrowleft\}^2$	"Narrow Beam"	Slow
2G			✓	$\{\rightarrow, \circlearrowleft\}^2$	Faster	Efficiency
3G			✓	$\{\circlearrowleft\}^2$	Faster	Efficiency, $
4G			×	$\circlearrowleft$	Less Moving Parts	More scattering
H*CT*		3G/4G	3G/4G	3G/4G	Fast 3D	Bit more expensive
MR*CT*			✓	$\{\circlearrowleft\}^2$	Fast 3D	Expensive
EB*CT*			×	None	Time Resolution	Expensive

✓*: Narrow Beam (implicitly collimated), H:Helical, MR:Multiple Row, EB:Electron Beam.

Second Generation Second-generation scanners improve upon this by introducing a linear detector array (Fig. 2.7a), such that one "fan beam" is measured simultaneously. Still, the fan beam is not wide enough to cover the full field of view. Initially, the opening angle of the fan beam was about 10°, consequently linear motion of the source-detector pair was still needed. The detector array greatly reduced the acquisition time at the cost of abandoning the narrow beam geometry, such that detector collimation was needed. By design, detector collimation reduces efficiency, such that for a given dose and identical projection measurements, a first-generation scanner would yield a better image. However, the win in reduced acquisition time greatly outweighs the drawbacks of reduced efficiency.

Third Generation Lots of scanners manufactured today utilize the general principle of third-generation *CT* scanners. Compared to second generation scanners, the opening angle of the fan beam grew to 40°–60° (see Fig. 2.7c). This means that linear motion is no longer necessary, as the fan beam can cover the full field of view. Similar to the leap from first to second-generation scanners, the acquisition time is greatly reduced, but again the efficiency of the detectors is lowered. This is necessarily the case, as without any linear motion the detector array must be very densely packed to obtain a sufficient number of samples per projection. This requires the detectors to be very small, leading to some loss in efficiency. Along with the number of detectors also increases the price of the system.

Fourth Generation In terms of image quality and acquisition time, fourth-generation scanners do not have any advantages over third-generation scanners. The difference lies in construction: As shown in Fig. 2.7d, fourth-generation scanners utilize a stationary (i.e. non-rotating)



(a) First-Generation

(b) Second-Generation

(c) Third-Generation

(d) Fourth-Generation

Figure 2.7: Illustration of the principles of different scanner generations. The different generations are discussed in detail in the text.

360° detector array, with only the X-Ray source rotating. Since the detectors need to acquire signals from many positions, they can not be collimated. Along with the possibility for physically larger detectors, this increases the detection efficiency as compared to third-generation scanners. However, the interference of scattering events does not allow better image quality. Such systems may be made more compact by positioning the source outside of the detector ring. To allow the detectors to "see" the source, the detector ring may experience out-of-plane nutation out of plane. Another possibility is to introduce gaps between the detectors to allow the X-Rays to pass through. We proceed by discussing classes of scanners with distinct characteristics, however, we do not assign a particular generation number.

Helical *CT* With the previously discussed scanner generations, exactly one slice of the body can be reconstructed during one acquisition run. Typically, the slice thickness ranges from 2 mm to 5 mm, and in order to acquire three-dimensional datasets, one would have to move the patient with respect to the detector ring by this distance. Clearly, this has the problem of being time consuming and prone to motion or misalignment artifacts. The obvious solution to this problem would be to slide the patient through the tube whilst it is rotating around the patient, acquiring data continuously. This is exactly the idea of helical *CT* scanners, which otherwise do not differ from third or fourth generation scanners. With this, it is possible to acquire full three-dimensional torso scans in around 30 s. Today, most systems are capable of helical acquisition, as they are only marginally more complicated in terms of hardware, and the reconstruction problem (whilst seemingly much more complicated) can be solved with simple interpolation techniques.

Multiple Row Detector (Cone Beam) *CT* It is natural to extend the detector array into the second dimension to get a detector matrix. Analogously to the detector array, the fan beam is extended into the third dimensions, such that multiple (these days up to 320) one-dimensional projections are acquired simultaneously. In some scanners, the array may be as high as it is wide, resulting in what is called *cone beam* geometry. If there is adequate distance between the X-Ray source and the detectors, approximately parallel planes can be imaged simultaneously. Combining this with helical scanning allows to acquire full three-dimensional datasets to be acquired in the order of seconds, with reduced dose compared to conventional helical systems.

Electron Beam *CT* Up until now, all the discussed scanner geometries relied on one X-Ray source. To get a full data set of one slice, this source (possibly also the detector array) has to be rotated around the patient. Due to the heavy construction, one rotation of the X-Ray tube along with the detector array or matrix usually takes around one second (although, as discussed above, multiple slices may be acquired during this time). This severely limits the ability to image (even a single slice) with high temporal resolution, e.g. for cardiac imaging or for tracing a contrast agent. For this purpose, electron beam scanners have been developed. These scanners produce the X-Rays by an electron beam steered by electromagnets, which hits a stationary tungsten anode ring. The resulting X-Rays are detected by a stationary detector ring. Since the electron beam can be steered very quickly (compared to the mechanical rotation of the tube-detector assembly), single-slice imaging can be done with a temporal resolution of about 50 ms.

Dual Source *CT* Dual (or multiple) energy *CT* can be easily acquired with traditional scanner geometries by simply running multiple scans or pulsing different tube voltages. However, this increases scan time and comes with other technical challenges, such that it is usually not done in clinical practice. To overcome these issues, dual source *CT* systems have been introduced, which operate with two physical tubes that allow different tube voltages. Multi-energy imaging is desirable because of diagnostic benefits: For instance, the composition of bones in the diagnosis of osteoporosis can be determined by such studies.

3 Image Formation

> Gradually he can see the reflections of people and things in water and then later see the people and things themselves.
>
> Plato, *Republic, VII, Allegory of the Cave*

Contents

In the previous chapter we discussed the underlying physical principles as well as how these principles can concretely be used in implementing scanners. Here, we want to investigate the mathematics of image acquisition and specifically reconstruction. Recall that in X-Ray *CT*, the fundamental measurement is a line integral of the linear attenuation coefficient $\mu(x, E)$. Specifically, assuming narrow beam geometry, we can relate the measured intensity I_d at the detector to the source spectrum $S_0(E)$ with

$$I_d = \int_0^{E_{\max}} S_0(\epsilon)\epsilon \exp\left(-\int_0^d \mu(\chi, \epsilon)\, \mathrm{d}\chi\right) \mathrm{d}\epsilon. \tag{3.1}$$

The above equation is in general intractable because of the energy dependence of the linear attenuation coefficient. Therefore, the effective energy $\bar{E}$ is introduced, see Section 2.2.3.4,

modifying Eq. (3.1) to

$$I_d = I_0 \exp\left(-\int_0^d \mu(\chi, \bar{E})\, \mathrm{d}\chi\right). \tag{3.2}$$

Finally, we can rearrange this to yield the line integral of the linear attenuation

$$g_d = -\log\left(\frac{I_d}{I_0}\right) = \int_0^d \mu(\chi, \bar{E})\, \mathrm{d}\chi. \tag{3.3}$$

In practice, the reference intensity I_0 can be measured for each detector in a calibration step.

3.1 Forward Problem — The Radon Transform

We have seen how a measurement in *CT* corresponds to a line integral of $\mu(x, \bar{E})$ along some path. However, what we desire is not the line integral, but the function $\mu(x, \bar{E})$ itself for any point x. Thus, we have to ask the question whether it is possible to reconstruct $\mu(x, \bar{E})$ given "all" its line integrals. Fortunately, this question has been answered by Johann Radon approximately 50 years prior to the first practical experiments with the predecessors of modern *CT*.

In "Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten" [77], Radon showed that a function $f : \mathbb{R}^2 \to \mathbb{R}$ is uniquely determined by its integrals along all lines. Specifically, let

$$\Theta = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \tag{3.4}$$

and

$$\Theta = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}. \tag{3.5}$$

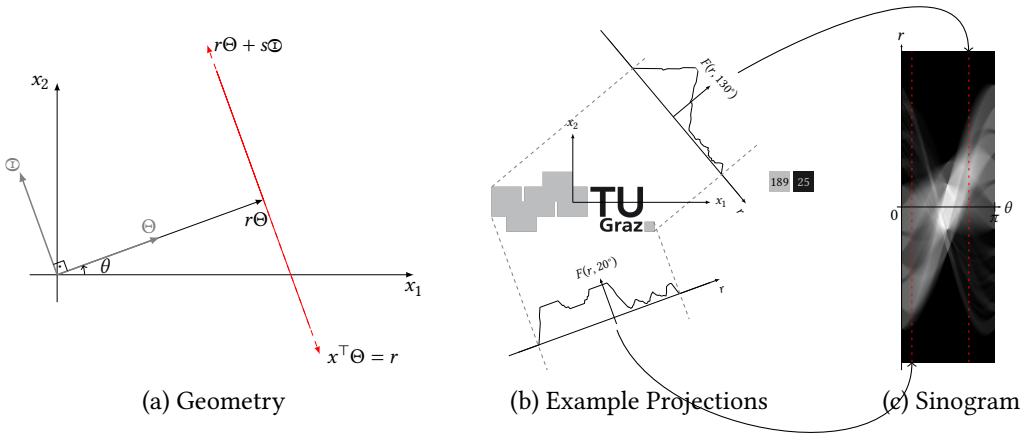
Then, f is uniquely determined by

$$\begin{aligned} F : \mathbb{R} \times [0, \pi] &\to \mathbb{R}, \\ (r, \theta) &\mapsto \int_{-\infty}^{\infty} f(r\Theta + s\Theta)\, \mathrm{d}s. \end{aligned} \tag{3.6}$$

This is true under mild assumptions about f, which are in general fulfilled in *CT*. Specifically, we require

1. f continuous,
2. $\displaystyle\int_{\mathbb{R}^2} \frac{|f(x)|}{\|x\|_2}\, \mathrm{d}x$ converges, and
3. $\displaystyle\lim_{r\to\infty} \int_0^{2\pi} f(x + r\Theta)\, \mathrm{d}\theta = 0 \;\forall x \in \mathbb{R}^2$.

Since the human body is finite in extent with finite μ and the linear attenuation coefficient of air is 0, the points are trivially fulfilled. In what follows, we assume that the image of f is $\mathbb{R}^+$.



(a) Geometry (b) Example Projections (c) Sinogram

Figure 3.1: In a, we show the interpretation of Θ, Θ, r, s. Note that $\{x \in \mathbb{R}^2 : x^\top\Theta = r\}$ and $\{r\Theta + s\Theta\}, s \in \mathbb{R}$ describe the same line in the plane. In b, the projections $F(r, 130°)$ and $F(r, 20°)$ of some function $f(x)$ are visualized, with the corresponding vertical lines in the sinogram c. For visualization purposes, we only show $\text{supp}(f)$, and the color-bar is shown on the right.

We call F the Radon transformed of f, and refer to $F(r, \theta')$ as a *projection* at a fixed angle θ'. To denote the transformation, we write $F(r, \theta) = (\mathscr{R}_2 f)(r, \theta)$. Note that we may write Eq. (3.6) as an integral over the two-dimensional Euclidean space as

$$(\mathscr{R}_2 f)(r, \theta) = \int_{\mathbb{R}^2} f(x) \underline{\infty}_{\{\xi \in \mathbb{R}^2 \,:\, \xi^\top\Theta = r\}}(x) \, \mathrm{d}x, \tag{3.7}$$

where we used the masking property of the delta distribution

$$\begin{aligned} \underline{\infty}_{\mathscr{I}} : \mathbb{R}^2 &\to \{0, \infty\}, \\ x &\mapsto \begin{cases} \infty & \text{if } x \in \mathscr{I}, \\ 0 & \text{else.} \end{cases} \end{aligned} \tag{3.8}$$

We can draw $F(r, \theta)$ in rectilinear coordinates to produce a *sinogram*. In scanner-geometry terms, we may think of r as the linear position of the detector, and we may think of θ as the angle of rotation of the gantry with respect to the fixed patient-coordinate system. To illustrate the ideas that were used above, we show an example in Fig. 3.1.

We want to note that, by construction, we are only considering integrals along parallel rays in each projection. This is theoretically only fulfilled in first-generations scanners, which are no longer used in medical applications. However, the following considerations can also be applied to fan beam geometries by adapting them accordingly. As such, going through the principles of the simplest geometries is a useful exercise.

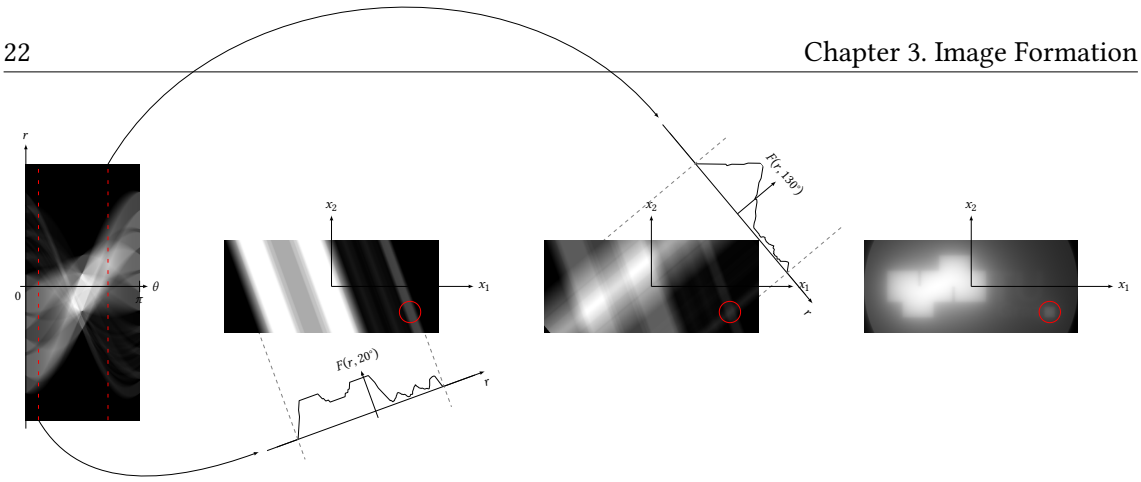


Figure 3.2: Illustration of the *SBP* algorithm: For every angle θ' in the sinogram, we smear $F(r, \theta')$ across the image. The red dot indicates the successive build-up of a distinct feature in the reconstruction process, and the reconstructed image is shown on the right.

3.2 The Inverse Problem

In the previous section, we specified the forward model by fixing the geometry and introducing some notation. Further, we hinted at the fact that Radon proved that any function can be uniquely reconstructed given its Radon transform. In this section, we will discuss some concrete reconstruction algorithms. Loosely speaking, the reconstruction algorithms can be divided into two categories: "analytic" and "algebraic". Analytic reconstruction essentially aims to reconstruct f from Eq. (3.6) or Eq. (3.7), while algebraic reconstruction techniques aim to solve the linear equation systems that arise in practice. In other words, analytic reconstruction solves the continuous model, while algebraic reconstruction solves the fully discretized model. We will point out correspondences between the two when they arise.

3.2.1 Simple Back-Projection

Considering the forward operation described in Section 3.1, one possible reconstruction method is tempting: What happens when we conceptually "reverse" the acquisition process? Specifically, consider a projection $F(r, \theta')$ at some angle $\theta' \in [0, \pi]$. As previously described, for a given $r' \in \mathbb{R}$, $F(r', \theta')$ is given by the integral of the desired function f along the line $x^\top \Theta' = r'$, where $\Theta' = (\cos\theta', \sin\theta')^\top$. Intuitively, the reverse of this would be to smear $F(r', \theta')$ along $x^\top \Theta' = r'$, i.e. to "back-project" it. If we do this for all r and θ, we get the *back-projected* image.

Mathematically, let $F = \mathscr{R}_2 f$, then the back-projected image $b : \mathbb{R}^2 \to \mathbb{R}$ is

$$b(x) = \int_0^\pi F(x^\top \Theta, \theta)\, \mathrm{d}\theta. \tag{3.9}$$

Notice that if $\operatorname{supp}(f) \neq \emptyset$ it follows that $b(x) > 0\ \forall x$. It is immediately clear that this is a drawback if we consider some $x' \notin \operatorname{supp}(f)$. We can conclude that b is in general not a faithful reconstruction of f. We illustrate Simple Back-Projection (SBP) for a toy example in Fig. 3.2.

Let us examine the problem of *SBP* in detail. By substituting Eq. (3.7) into Eq. (3.9), we find

that

$$b(x) = \int_0^\pi \int_{\mathbb{R}^2} f(\chi) \underline{\infty}_{\{\xi \in \mathbb{R}^2 \,:\, \xi^\top \Theta = x^\top \Theta\}}(\chi) \, \mathrm{d}\chi \, \mathrm{d}\theta. \tag{3.10}$$

We change the order of integration and note the equivalence $\{\xi \in \mathbb{R}^2 : \xi^\top \Theta = x^\top \Theta\} = \{\xi \in \mathbb{R}^2 : (\xi - x)^\top \Theta = 0\}$ to yield

$$b(x) = \int_{\mathbb{R}^2} f(\chi) \left(\int_0^\pi \underline{\infty}_{\{\xi \in \mathbb{R}^2 \,:\, (\xi - x)^\top \Theta = 0\}}(\chi) \, \mathrm{d}\theta \right) \mathrm{d}\chi. \tag{3.11}$$

Let ϕ denote the angle between $(\xi - x)$ and the x_1-axis, then $\{\xi \in \mathbb{R}^2 : (\xi - x)^\top \Theta = 0\} = \{\xi \in \mathbb{R}^2 : \|\xi - x\| \cos(\phi - \theta) = 0\}$. We now use

$$\underline{\infty}_{\{0\}}(g(x)) = \sum_{\{x_i \,:\, g(x_i) = 0\}} \left| \frac{\mathrm{d}g}{\mathrm{d}x}(x_i) \right|^{-1} \underline{\infty}_{\{x_i\}}(x), \tag{3.12}$$

which is a well known identity in δ-calculus [8], and note that, assuming $\xi \neq x$, $\|\xi - x\| \cos(\phi - \theta) = 0 \Leftrightarrow \phi - \theta = \pm\frac{\pi}{2}$. We therefore obtain

$$b(x) = \int_{\mathbb{R}^2} f(\chi) \left(\int_0^\pi \frac{\underline{\infty}_{\{\pm\frac{\pi}{2}\}}(\theta)}{\|\chi - x\| \, |\sin \pm\frac{\pi}{2}|} \, \mathrm{d}\theta \right) \mathrm{d}\chi. \tag{3.13}$$

Since the delta distribution integrates to 1 and the denominator is independent of θ, this is easily simplified to

$$b(x) = \int_{\mathbb{R}^2} f(\chi) \frac{1}{\|\chi - x\|} \, \mathrm{d}\chi. \tag{3.14}$$

Clearly, Eq. (3.14) is a convolution. Thus, one may equivalently write

$$b(x) = f(x) * h(x), \tag{3.15}$$

with the point spread function $h(x) = \frac{1}{\|x\|}$. In general, this blurring makes *SBP* useless, as too much diagnostic value is lost and more precise reconstruction algorithms with marginally higher computational cost exist.

However, one may ask the obvious question: Since we know the blur function, what if we simply deconvolve b with h, which can easily be done in the Fourier domain? In fact, f can be faithfully reconstructed that way, and this is mathematically equivalent with one of the most popular reconstruction algorithms, namely *Filtered Back-Projection (FBP)* (although *FBP* has practical benefits). In what follows, we will introduce the famous *Fourier Slice Theorem*, which gives rise to a family of Fourier-based reconstruction algorithms.

3.2.2 Fourier Slice Theorem

The Fourier Slice Theorem will allow us to derive an "inverse Radon transform". Consider the projection $F(r, 0)$ of f, i.e.

$$F(r, 0) = \int_{-\infty}^{\infty} f(r, x_2)\, \mathrm{d}x_2. \tag{3.16}$$

The one-dimensional Fourier transform of $F(r, 0)$ with respect to r is

$$(\mathscr{F}_1 F)(\xi, 0) = \int_{-\infty}^{\infty} F(\chi, 0) \exp\left(-\mathrm{i}2\pi\xi\chi\right) \mathrm{d}\chi \tag{3.17}$$

with the imaginary unit i. Substituting Eq. (3.16) yields

$$\begin{aligned}(\mathscr{F}_1 F)(\xi, 0) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2)\, \mathrm{d}x_2 \exp\left(-\mathrm{i}2\pi\xi x_1\right) \mathrm{d}x_1 \\ &= \int_{\mathbb{R}^2} f(x) \exp\left(-\mathrm{i}2\pi(\xi x_1 + 0x_2)\right) \mathrm{d}x = (\mathscr{F}_2 f)(\xi, 0)\end{aligned} \tag{3.18}$$

In other words, the one-dimensional Fourier transform of the projection of f at $\theta = 0$ maps to the same values as the two-dimensional Fourier transform of f along the horizontal components.

If we consider $F(r, 0)$ as the projection onto the x_1' axis of some rotated coordinate system, the computation above holds. In general, the two-dimensional Fourier transform of a function f rotated by some angle α is also rotated by α with respect to $\mathscr{F}_2 f$. Therefore, we proved the famous Fourier Slice Theorem, which we summarize as follows:

> Let $f : \mathbb{R}^2 \to \mathbb{R}$, and let $\mathscr{F}_2 f$ be its two-dimensional Fourier transform. Further, let $F = \mathscr{R}_2 f$ be its two-dimensional Radon transform, with $\mathscr{F}_1 F$ its one-dimensional Fourier transform with respect to the affine parameter. Then, $(\mathscr{F}_1 F)(\cdot, \theta)$ describes the values of $\mathscr{F}_2 f$ on the radial line at angle θ. Thus,
>
> $$(\mathscr{F}_1 F)(\rho, \theta) = (\mathscr{F}_2 f)(\rho\Theta). \tag{3.19}$$

3.2.3 Direct Fourier Inversion

Knowing the Fourier Slice Theorem, another obvious reconstruction algorithm arises: We can populate the $\mathscr{F}_2 f$ along radial lines of angle θ with $(\mathscr{F}_1 F)(\rho, \theta)$, and then reconstruct f by

$$f = \mathscr{F}_2^{-1}(\mathscr{F}_2 f). \tag{3.20}$$

This is schematically shown in Fig. 3.3. While straight forward in theory, this approach is usually not used in practice.

In practice, scanners acquire a finite number of N projections $\{F(r, \theta_1), \ldots, F(r, \theta_N)\}$ during one scanner rotation. These measurements fill the Fourier space in radial lines, such that we

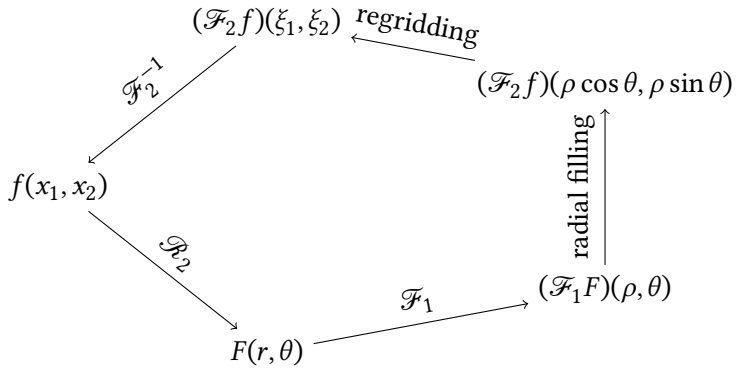


Figure 3.3: Schematic of the direct Fourier method for inverting the Radon transform. The error that arises during interpolation in the regridding step is what prevents this method from being used in practice.

need to interpolate the Cartesian grid prior to performing the inverse Fourier transform. This process is called *regridding* and is problematic in practice, since the distance between the radial streaks increases with the spatial frequency. As such, the interpolation becomes less precise for higher frequencies, which encode the details in the image. There exist schemes for non-uniform sampling in the Radon space to alleviate this issue, but they are not widely used and we do not discuss them here. We refer the interested reader to [65] for further discussion.

3.2.4 Filtered Back-Projection

Fortunately, we can further modify the ideas of the direct Fourier inversion method to yield a reconstruction scheme that is very useful in practice. We derive the basis equation of *FBP* by considering the inverse Fourier transform of the image, i.e.

$$f(x) = \int_{\mathbb{R}^2} (\mathscr{F}_2 f)(\xi) \exp\left(\mathrm{i}2\pi x^\top \xi\right) \mathrm{d}\xi. \tag{3.21}$$

We introduce the polar coordinates $\rho\Theta = \xi$, $\mathrm{d}\xi = \rho \, \mathrm{d}\rho \, \mathrm{d}\theta$, such that

$$f(x) = \int_0^{2\pi} \int_0^{\infty} (\mathscr{F}_2 f)(\rho\Theta) \exp\left(\mathrm{i}2\pi\rho x^\top \Theta\right) \rho \, \mathrm{d}\rho \, \mathrm{d}\theta, \tag{3.22}$$

and note that we can equivalently scan the plane by (a very rigorous proof can be found in [12])

$$f(x) = \int_0^{\pi} \int_{-\infty}^{\infty} (\mathscr{F}_2 f)(\rho\Theta) \exp\left(\mathrm{i}2\pi\rho x^\top \Theta\right) |\rho| \, \mathrm{d}\rho \, \mathrm{d}\theta. \tag{3.23}$$

With the Projection Slice Theorem Eq. (3.19), we may write

$$f(x) = \int_0^{\pi} \int_{-\infty}^{\infty} |\rho| (\mathscr{F}_1 F)(\rho, \theta) \exp\left(\mathrm{i}2\pi\rho x^\top \Theta\right) \mathrm{d}\rho \, \mathrm{d}\theta, \tag{3.24}$$

or by noting that $x^\top \Theta$ is nothing else than the "detector position" r,

$$f(x) = \int_0^\pi \underbrace{\left[\int_{-\infty}^{\infty} |\rho| (\mathscr{F}_1 F)(\rho, \theta) \exp(\mathrm{i}2\pi\rho r) \, \mathrm{d}\rho \right]}_{\tilde{F}(r,\theta)} \mathrm{d}\theta. \tag{3.25}$$

3.2.4.1 The Filtered Projections

Let us now pay closer attention to the term in the square brackets, which we detail

$$\tilde{F}(r, \theta) = \int_{-\infty}^{\infty} |\rho| (\mathscr{F}_1 F)(\rho, \theta) \exp(\mathrm{i}2\pi\rho r) \, \mathrm{d}\rho. \tag{3.26}$$

This is exactly the one-dimensional inverse Fourier transform, weighted by $|\rho|$. In other words, if we disregard this factor, we would have $\tilde{F} = \mathscr{F}_1^{-1} \mathscr{F}_1 F = F$, which are the original projections. By the famous convolution theorem, a multiplication in the Fourier domain corresponds to a convolution in the spatial domain. Therefore, we may interpret $|\rho|$ as a filter acting on the projections F, i.e. $\tilde{F}$ are filtered (read: convolved with some filter response) projections.

In the Fourier domain, the influence of $|\rho|$ is relatively straight-forward: Since the "radius" ρ essentially describes the frequency, we can immediately conclude that a multiplication with $|\rho|$ in the Fourier domain is a high-pass filter. On the other hand, since $|\rho|$ is not square-integrable, we can not simply calculate its inverse Fourier transform to get the spatial filter. However, we can detail the spatial filter by a limit process. As an example, we may consider

$$\left(\mathscr{F}_1 \left\{ \frac{\epsilon^2 - (2\pi r)^2}{\left(\epsilon^2 + (2\pi r)^2\right)^2} \right\} \right)(\rho) = |\rho| \exp(-\epsilon|\rho|), \tag{3.27}$$

where $\lim_{\epsilon \to 0} |\rho| \exp(-\epsilon|\rho|) = |\rho|$. We show these functions in Fig. 3.4. In summary, in *FBP*, we filter the projections by multiplying them with $|\rho|$ in the frequency domain.

3.2.4.2 The Back-Projection

Let us now consider the outer integral in Eq. (3.25), namely

$$f(x) = \int_0^\pi \tilde{F}(x^\top \Theta, \theta) \, \mathrm{d}\theta. \tag{3.28}$$

This is exactly Eq. (3.9), where we replaced the projections F with their filtered counterparts $\tilde{F}$. The intuition is the same as discussed in Section 3.2.1. That is, in the image $f(x)$ we smear $\tilde{F}(r, \theta)$ along the line $x^\top \Theta = r$. Summing this over all $\theta \in [0, \pi]$ (i.e. carrying out the integration) finally yields the reconstructed image $f(x)$.

Thus, we summarize *FBP* as a reconstruction algorithm in three steps:

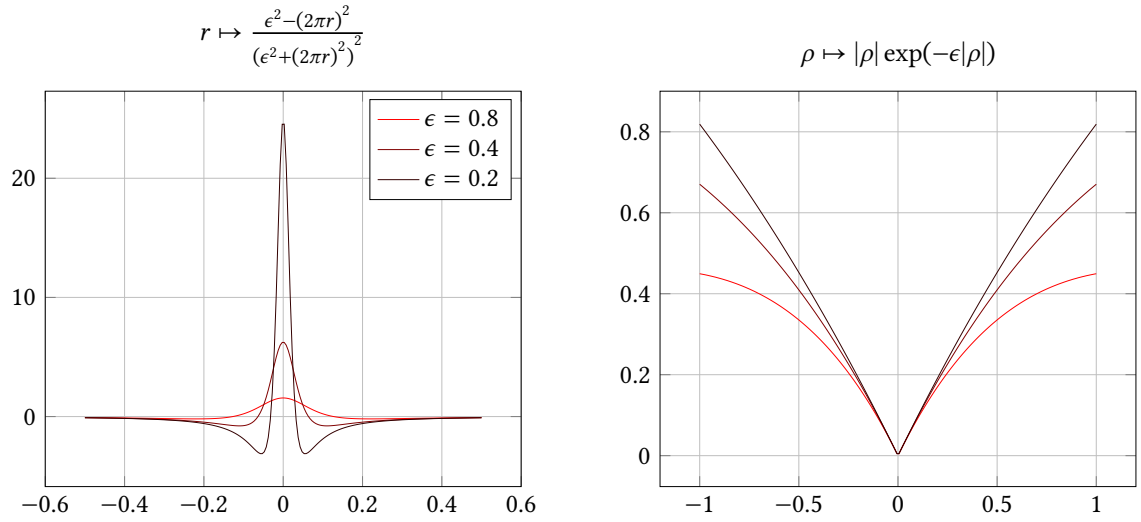


Figure 3.4: Approximations for spatial convolution filters mimicking a multiplication with $|\rho|$ in the Fourier domain.

Let $F = \mathscr{R}_2 f$ be the Radon transform of a function $f : \mathbb{R}^2 \to R$. Then, the *FBP* algorithm for reconstructing f is as follows:

1. Calculate the one-dimensional Fourier transform $(\mathscr{F}_1 F)(\rho, \theta)$ of $F(r, \theta)$.
2. High-pass filter $(\mathscr{F}_1 F)(\rho, \theta)$ with $|\rho|$ and compute the inverse Fourier transform
$$\tilde{F}(r, \theta) = \mathscr{F}_1^{-1}\{|\rho|(\mathscr{F}_1 F)(\rho, \theta)\}.$$
3. Back-project the filtered Radon transform $\tilde{F}$ onto the image by
$$f(x) = \int_0^\pi \tilde{F}(x^\top \Theta, \theta)\, \mathrm{d}\theta.$$

We show reconstruction of our toy example in Fig. 3.5, where we see that the image is faithfully reconstructed.

In the discussion about the *SBP* algorithm, we hinted at the fact that one may also deconvolve the resulting image with the known blur function. The corresponding reconstruction algorithm is sometimes referred to as *filtered layergram*. *FBP* is mathematically equivalent to this, but it has a practical advantage: Since the projections can be filtered independently, the reconstruction may start as soon as the first projection is acquired. Moreover, looking at Fig. 3.4 we see that the spatial approximations of $|\rho|$ have a small support. Therefore, in practice it is often advantageous to completely circumvent all frequency domain calculations by implementing the spatial convolution rater than the Fourier domain multiplication. The spatial kernel may also be calculated by

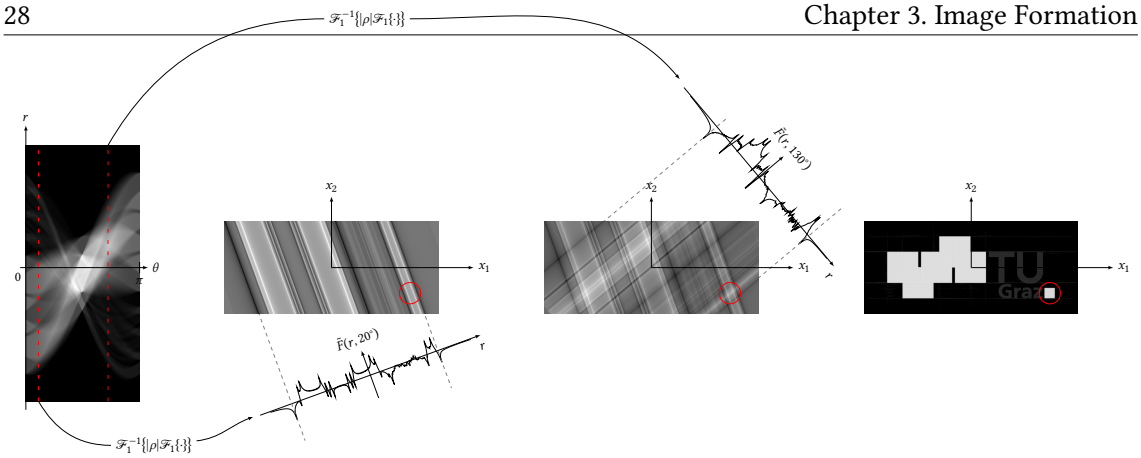


Figure 3.5: In the *FBP* algorithm, the filtered projections $\tilde{F}(r, \theta)$ are smeared across the image. The red circles indicate the successive build-up of a distinct feature in the image, and on the right we show the final reconstruction.

windowing $|\rho|$, or by simply choosing an appropriate kernel that may not be motivated by the frequency domain multiplication [78].

3.3 Algebraic Reconstruction

The reconstruction algorithms that were discussed in the previous sections are fast and derived in a rigorous framework. While this may seem desired at first glance, it is actually a big weakness: Due to wrong initial assumptions about the model, specifically that the X-Ray source is a monochromatic zero-width beam, these algorithms introduce typical artifacts in the reconstruction (see Section 3.4). Accounting for the energy dependence of the linear attenuation coefficient μ is intractable mathematically and practically, and it is very hard to "inject" prior knowledge into the analytic reconstruction algorithms.

Algebraic (more specifically, iterative algebraic) methods allow for easy re-weighting of rays, and allow prior knowledge to be considered in the reconstruction. This comes at the cost of much more computational expense, which is why typically analytic reconstruction dominated *CT* historically. However, the increase in computational power over the recent years has caused a gradual shift towards iterative methods. In what follows, we want to detail the discretization of the model, and present typical algorithms for solving the arising linear system of equations. We discuss linear inverse problems in the general sense, but put particular focus on the practical problems that arise specifically in *CT* reconstruction.

3.3.1 Discretized Model

The analytic Fourier-based reconstruction algorithms are very principled in theory. However, in practice the acquisition system is inherently discrete by design of the detector elements, and we store and view the reconstructed image on a discretized grid of picture elements. Further, the integration along a line can not be realized practically, as the X-Ray beam will always have some

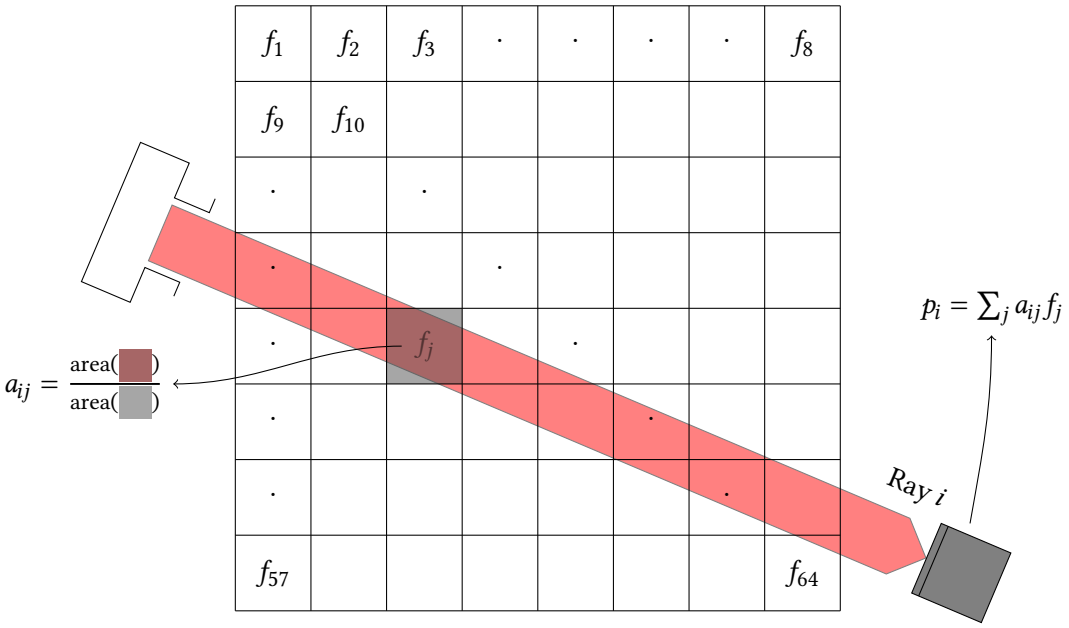


Figure 3.6: The fully discretized model in *CT* reconstruction, where the weights a_{ij} of the forward operator A are given by the area of the intersection of the i-th beam with the j-th pixel.

"width" to it.

We represent the discretized image f as an N-dimensional vector, which corresponds to the *flattened* image, i.e. is numbered in row-major order. The projections p_i through f are easily modeled by introducing a weighting factor for each ray and pixel. There exist many ways of defining the weighting, such as length of the ray in the pixel (assuming a zero-width ray) or the area of the intersection of the ray with the pixel. In this case, for any pixel index j, the weighting factor a_{ij} is, loosely speaking, the cross section of f_j with the i-th ray, divided by the total area of f_j. We show an example for the area integration framework schematically in Fig. 3.6. The evaluation of the forward operator is a very costly operation computationally, and much work has gone into efficiently computing it. As an example, we refer the reader to [39] for a look-up table-based area integration approach.

With Fig. 3.6, we may now enumerate the equations as

$$\begin{aligned} a_{11}f_1 + a_{12}f_2 + \cdots + a_{1N}f_N &= p_1 \\ a_{21}f_1 + a_{22}f_2 + \cdots + a_{2N}f_N &= p_2 \\ &\vdots \\ a_{M1}f_1 + a_{M2}f_2 + \cdots + a_{MN}f_N &= p_M, \end{aligned} \tag{3.29}$$

or equivalently

$$p = Af, \tag{3.30}$$

with the measurements $p \in \mathbb{R}^M$, the design or system matrix $A \in \mathbb{R}^{M\times N}$ and the image $f \in \mathbb{R}^N$.

Table 3.1: Comparison of analytic reconstruction and algebraic reconstruction.

Reconstruction	Model	Speed	Flexibility
Analytic	continuous	fast	None
Algebraic	discrete	slower	High

Throughout this work, we assume that A is known, and is appropriate for the problem. We now note the equivalence with the continuous Radon transform by

$$\begin{aligned} p &= Af \\ &\Updownarrow \\ F &= \mathscr{R}_2 f_{\text{cont}}, \end{aligned} \tag{3.31}$$

i.e. p represents the sinogram in the Radon space and f holds the image values in the image domain. The matrix A represents the linear map from the image space to the Radon space.

For example, assume we want to reconstruct an image $f \in \mathbb{R}^N$ where $N = 512 \times 512 = 262\,144$. Further, assume acquisition of $M = 1000 \times 500 = 500\,000$ rays, whereby $N_D = 500$ detectors acquire $N_P = 1000$ projection directions. Then, the system matrix $A \in \mathbb{R}^{262\,144 \times 500\,000}$ maps the image f to the sinogram p. We already see a practical problem arising: Storing the system matrix A naively with 32 bit floating-point numbers would require approximately 524 GB of storage. Fortunately, as seen in Fig. 3.6, the i-th ray does not intersect most pixels at all. In other words, most entries a_{ij} in A are 0, so we call A "sparse" and can store it efficiently.

The size and structure of A is still a problem. Specifically, it is large without "simple" structure (to allow special inversion algorithms), such that in general we need *iterative methods* to solve Eq. (3.30). However, this allows to easily deal with irregularities in the measurement data, or to incorporate prior knowledge into the reconstruction problem. We summarize the advantages and disadvantages of algebraic reconstruction over analytic reconstruction in Table 3.1.

3.3.2 Solving the Linear System

Let us consider Eq. (3.30) in more detail. In practice, the system will be overdetermined, since more projections than pixels are acquired. Further, the measurements are noisy — that is, in reality we only have access to noisy projection data $p = Af + v$. A well known solution to overdetermined, noisy systems is the least-squares solution

$$f^*_{\text{LS}} = \arg\min_f \|Af - p\|_2^2, \tag{3.32}$$

which can be easily solved in closed form (disregarding the practical problem of actually *calculating* it) as

$$f^*_{\text{LS}} = (A^\top A)^{-1} A^\top p = A^\dagger p. \tag{3.33}$$

Here, $A^\dagger = (A^\top A)^{-1} A^\top$ is the *Moore-Penrose pseudo inverse* of A.

In Eq. (3.31), we noted how A may be the discretized Radon transform, mapping from image space into Radon space. The *adjoint* operation, i.e. mapping from Radon space to image space is given by $A^\top$. Thus, we may interpret $b = A^\top p$ as the "simple back-projection". Then it is clear that $(A^\top A)^{-1}$ "filters" the back-projected image. In this sense, $f^*_{\mathrm{LS}} = (A^\top A)^{-1} A^\top p = (A^\top A)^{-1} b$ may be interpreted as the filtered layergram algorithm. In the continuous discussion, we mentioned that the order of filtering and back-projection may be reversed. Similarly, in the discrete setting we may write

$$f^*_{\mathrm{LS}} = A^\top (AA^\top)^{-1} p, \tag{3.34}$$

where $(AA^\top)^{-1}$ is the linear frequency ramp filter, and $A^\top$ is the back-projection operator.

Although we got valuable insight into the relationship between the continuous and discrete equations, the direct inversion does not work in practice because of the size of the matrices that need to be inverted. In general, Eq. (3.30) is solved by iterative methods, which we will discuss in the following section.

3.3.3 Iterative Reconstruction

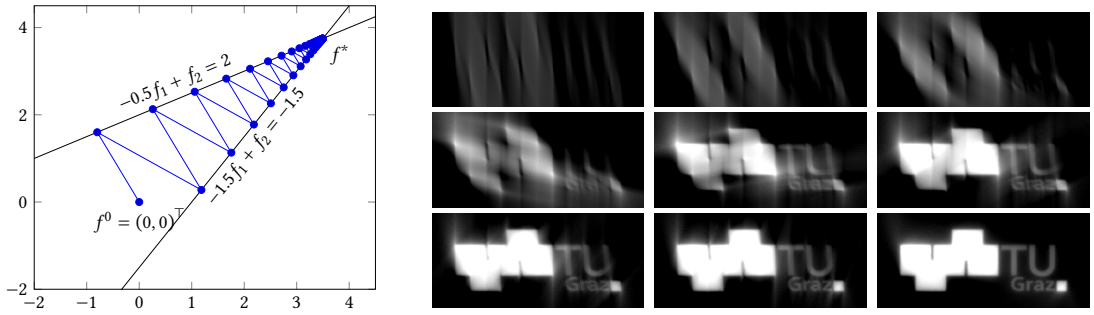
Iterative reconstruction techniques aim to solve linear problems of the form Eq. (3.30) without explicitly calculating the (generalized) inverse of A. The principle of all iterative reconstruction techniques can be summarized as follows:

1. Forward Projection: Given the current estimate f^*, compute the "expected" projections $\hat{p}$.
2. Correction: Compute the "error" $(\hat{p} - p)$ between the expected projections and the measured projections, and properly normalize it.
3. Back-Projection: Distribute the error back across all the pixels f^*_i that contributed to the difference.

The iterative algorithms that are typically used in *CT* reconstruction differ in the selection of pixels or beams that are considered simultaneously. For instance, the forward projection may consider all pixels simultaneously or one pixel after the other, or we may operate on a ray-by-ray basis and correct only the pixels that contribute to the projection of the current ray.

3.3.3.1 Kaczmarz Method

Maybe the conceptually simplest algebraic reconstruction algorithm is the Kaczmarz method or method of projections [48], sometimes also (ambiguously) referred to simply as Algebraic Reconstruction Technique (ART). It is of the ray-based family of reconstruction techniques, where the corrections are applied on a ray-by-ray basis. Specifically, the Kaczmarz iterations proceed by orthogonally projecting the images onto the hyperplanes defined by the projections of the individual rays, one after another.



(a) (b)

Figure 3.7: In a, we show the Kaczmarz iterations to solve a simple linear system, where the solution converges to $f^* = (3.5, 3.75)^\top$. Each iteration is an orthogonal projection onto the hyperplane corresponding to one row in the system. In b, a tomographic reconstruction is shown, where after one iteration (lower right) the image is satisfactorily reconstructed.

Mathematically, this reads

$$f \leftarrow f - \frac{A_i f - p_i}{\|A_i\|_2^2} A_i^\top, \quad \forall i = 1, \dots, M, \tag{3.35}$$

where $A_i = (a_{i1}, a_{i2}, \dots, a_{iN})$ is the i-th row of A. Note how in Eq. (3.35) $A_i f$ "forward projects" the image f, the result of which is then compared to the measured projection p. The resulting error is normalized by $\|A_i\|_2^2$, and subsequently "back projected" by $A_i^\top$. In this sense, Eq. (3.35) describes a cyclic gradient descent with adaptive step size $\frac{1}{\|A_i\|_2^2}$.

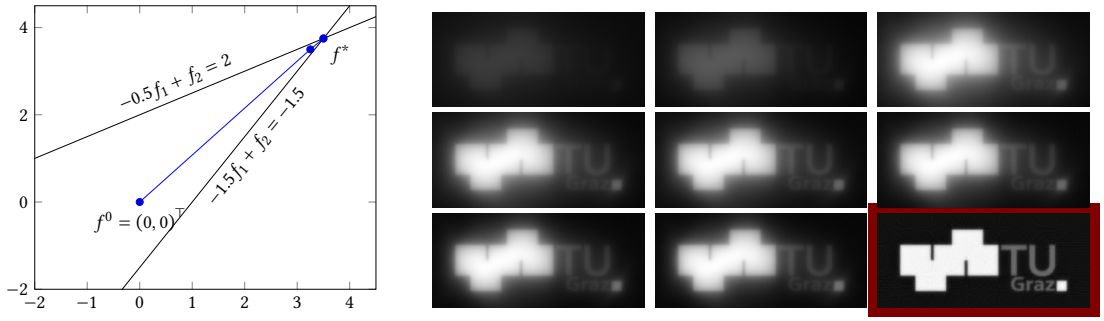
Typically we say that one iteration of the Kaczmarz algorithm is finished after all rows $A_i, i = 1, \dots, M$ have been considered once. Note that, since the systems arising in practical *CT* are usually strongly overdetermined (i.e. $M \gg N$), most often the algorithm gives satisfactory results in one iteration. In fact, practically the idea of an iteration is discarded and the "row index" i in Eq. (3.35) is often randomized. We show the *ART* algorithm graphically in Fig. 3.7.

3.3.3.2 Simultaneous Iterative Reconstruction Technique

An obvious modification to the *ART* algorithm is to consider not one, but all rays *simultaneously* during one iteration. The Simultaneous Iterative Reconstruction Technique (SIRT) algorithm does just that: Instead of updating the image successively to minimize the difference to each individual projection, it accumulates (and properly normalizes) the updates of *all* rays during one iteration. This modifies Eq. (3.35) to

$$f \leftarrow f - CA^\top R(Af - p) \tag{3.36}$$

where C, R are diagonal matrices containing the inverse of the column- and row sums of A, that is $c_{jj} = \left(\sum_i a_{ij}\right)^{-1}$ and $r_{ii} = \left(\sum_j a_{ij}\right)^{-1}$. We show the iterations of the *SIRT* for the same problem



(a) (b)

Figure 3.8: For the toy example in a, after two *SIRT* iterations we converge to f^*. In a, we show the first 8 *SIRT* iterations ($f^0 = 0$), and the final reconstruction after 100 iterations (highlighted in red).

as in Fig. 3.7 in Fig. 3.8.

3.3.3.3 Simultaneous Algebraic Reconstruction Technique

Given *ART* and *SIRT*, it is natural to consider some "in-between" cases. In Simultaneous Algebraic Reconstruction Technique (SART), one considers all rays of a particular projection simultaneously. We may write this as

$$f \leftarrow f - C_V A_V^\top R(A_V f - p), \tag{3.37}$$

where $A_V \in \mathbb{R}^{v \times N}$ contains the rows of A that correspond to a certain "view" (i.e. rotation angle), and C_V is contains the inverse of the corresponding column sums.

Of course, the field of linear systems is well studied and a plethora of algorithms exist for solving Eq. (3.30). Other algorithms that are frequently used in *CT* are the Block Iterative Component Averaging (BICAV) [14], Ordered Subset Separable Quadratic Surrogates (OS-SQS) [50, 53], and Conjugate Gradient (CG) [31, 79].

3.4 Artifacts

In this section we will discuss typical artifacts in *CT* imaging. In the following discussion, we will take a broad definition of "artifact". Specifically, we define an artifact as any difference between the reconstruction and the measured function. Artifacts may therefore appear because of the simplifications of the physical model, the fact that we only sample the radon transform, noise in the measurements, and movement of the patient during acquisition.

3.4.1 Finite Beam Width

In the derivation of the analytic reconstruction algorithms, it was always assumed that we can can integrate f along lines of no width. Obviously, with real source-detector pairs this is not the case. Even if we assume continuous sampling along the affine parameter r, we would acquire the *strip integral*

$$(\mathscr{R}_2^w f)(r,\theta) = \int_{-\infty}^{\infty} w(u)(\mathscr{R}_2 f)(r-u,\theta)\,\mathrm{d}u, \tag{3.38}$$

where $w:\ \mathbb{R} \to \mathbb{R}^+$ is a weight function, sometimes called *beam profile*. It summarizes the effects of the finite-width source beam and detector pair.

3.4.1.1 Image Convolution

Shepp and Logan showed that the weighted Radon transform $(\mathscr{R}_2^w f)$ is the Radon transform of a convolved signal $f * k$ [89]. Specifically, $\mathscr{R}_2^w f = \mathscr{R}_2(f * k)$ where $k:\ \mathbb{R}^2 \to \mathbb{R}^+$ is the radial function

$$k(x) = -\frac{1}{\pi \|x\|}\partial_{\|x\|}\int_{\|x\|}^{\infty} \frac{w(u)u}{\sqrt{u^2 - \|x\|^2}}\,\mathrm{d}u. \tag{3.39}$$

Note that, if w has bounded support (which is a reasonable assumption in practice), then k also has finite support. The following is an example pair (w,k):

$$w(u) = \begin{cases} \frac{1}{2d} & \text{if } -d \le u \le d, \\ 0 & \text{else,} \end{cases} \qquad k(x) = \begin{cases} \frac{1}{2\pi d}\frac{1}{\sqrt{d^2-\|x\|^2}} & \text{if } 0 \le \|x\| \le d, \\ 0 & \text{else.} \end{cases} \tag{3.40}$$

We conclude that, even if we assume continuous sampling of $(\mathscr{R}_2^w f)(r,\theta)$ along both (r,θ) and a perfect reconstruction algorithm, the finite width acquisition with beam profile w allows only to reconstruct $(f * k)$.

3.4.1.2 Partial Volume Effect

The finite-width beam manifests itself also in another typical artifact, which is the *partial volume effect*. Let us recall the idealized fundamental X-Ray attenuation law

$$I_d = I_0 \exp\left(-(\mathscr{R}_2 f)(r',\theta)\right), \tag{3.41}$$

where I_d is the measured intensity at the detector that corresponds to the alignment (r',θ). With a finite-width beam, we can modify this as

$$\log\frac{I_d}{I_0} = \log\left(\int_{-\infty}^{\infty} w(u)\exp\left(-(\mathscr{R}_2 f)(r'-u,\theta)\right)\,\mathrm{d}u\right), \tag{3.42}$$

where clearly the measurement depends on f in a nonlinear fashion.

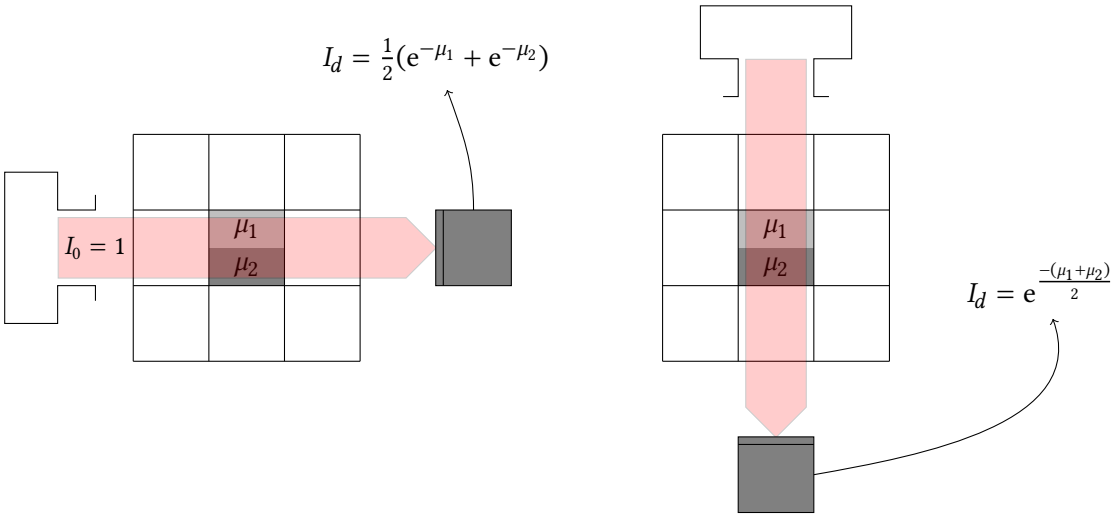


Figure 3.9: Inconsistent projections caused by the partial volume effect. We assume that $I_0 = 1$ and that the intensity is distributed equally between the sub-pixels on the left. Further, for simplicity we define the length of the pixels to be 1. We show the grid for visualization purposes but assume $\mu = 0$ everywhere outside the central pixel.

By Taylor expansion of Eq. (3.42), we can quantify the error of the linearization as

$$\log \frac{I_d}{I_0} = (\mathscr{R}_2^w f)(r', \theta) + \mathcal{O}\left(\int_{-\infty}^{\infty} w(u)\Big((\mathscr{R}_2 f)(r' - u, \theta) - (\mathscr{R}_2 f)(r', \theta)\Big)^2 \, \mathrm{d}u\right). \tag{3.43}$$

We see that the error depends on the (weighted) "variance" of $\mathscr{R}_2 f$ over the width of the beam. In concrete terms, the error is large if there are objects of drastically different linear attenuation coefficient within the beam width. In practice this is the case if bone or contrast agents partially intersect a pixel.

The partial volume effect may also occur "out of plane", i.e. if bone partially intersects the current slice from an adjacent slice. This is actually the more benign case, as here we would simply measure the wrong (i.e. not necessarily the mean of the two) attenuation coefficient at the affected pixels. If however the effect occurs "in plane", the projections will be inconsistent, such that they can not compensate each other properly outside of the pixel. We show the problem in Fig. 3.9. In such cases, the typical streaking artifacts occur.

3.4.2 View and Ray Sampling

Clearly, the assumption that we know $(\mathscr{R}_2^w f)(r, \theta)$ for all $r \in \mathbb{R}$ and $\theta \in [0, \pi]$ is misguided. In practice, the measurement process acquires samples from $\mathscr{R}_2^w f$ in both the affine and rotational argument. Assuming Δ_r is the distance between samples in the r direction, and similarly Δ_θ in the θ direction, we have access to the set of measurements

$$\{(\mathscr{R}_2^w f)(i_r \Delta_r, i_\theta \Delta_\theta) : \; i_r = -N_r, \dots, N_r, i_\theta = 0, \dots, N_\theta\}. \tag{3.44}$$

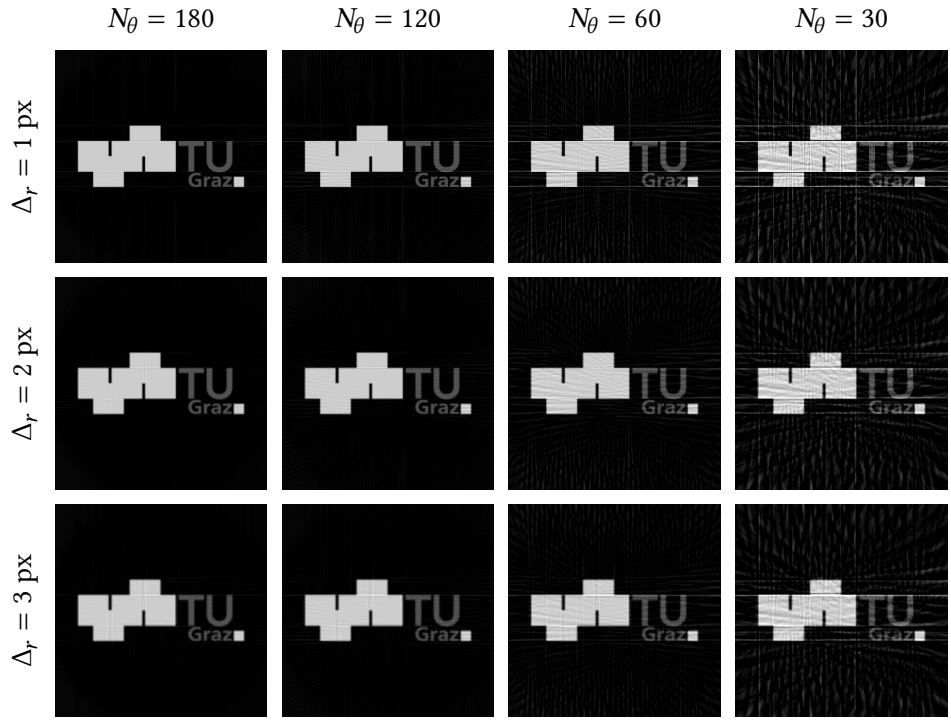


Figure 3.10: Artifacts in the reconstruction induced by sampling $\mathscr{R}_2^w f$ "sparsely". Note that we chose Δ_θ such that $\Delta_\theta N_\theta = 180°$.

Note that N_r is usually not critical, since we assume $f(x) = 0$ for $\|x\| > r_{\max}$ and usually $N_r \Delta_r > r_{\max}$.

Although it is possible to analytically calculate the influence of the sampling on the point spread function, it is beyond the scope of this work. We empirically show *FBP* reconstructions of images with varying Δ_r and Δ_θ in Fig. 3.10. Clearly, the view sampling results in the typical oscillations along lines tangent to sharp discontinuities. On the other hand, the ray sampling essentially low-pass filters the image, and also partial volume effects can be clearly seen.

3.4.3 Noise

To discuss the noise propagation in *CT*, we first show how it influences the sinogram, and then propagate this error through the filtered back-projection. Let

$$\bar{N}(r, \theta) = N_0 \exp\left(-\int_{-\infty}^{\infty} \mu(r\Theta + s\Theta) \, \mathrm{d}s\right) \tag{3.45}$$

be the number of detected photons of the detector at position r and angle θ, under the assumption of a perfect measurement. We denote with N_0 the number of photons expelled by the source, which we assume to be be a deterministic and known number, although in practice it also follows

some probability distribution. The line integral of the linear attenuation coefficient is therefore

$$\bar{g}(r,\theta) = \log \frac{N_0}{\bar{N}(r,\theta)} = \log N_0 - \log \bar{N}(r,\theta). \tag{3.46}$$

It is well known that the measured photons N actually follow a Poisson distribution, that is

$$\mathbb{P}(N(r,\theta) = c) = \frac{\left(\bar{N}(r,\theta)\right)^c \exp\left(-\bar{N}(r,\theta)\right)}{c!}, \tag{3.47}$$

with variance $\mathbb{V}$ and expected value $\mathbb{E}$

$$\mathbb{V}\left\{N(r,\theta)\right\} = \mathbb{E}\left\{N(r,\theta)\right\} = \bar{N}(r,\theta). \tag{3.48}$$

If we let g denote the measured measurement, then

$$\mathbb{E}\left\{g(r,\theta)\right\} = \mathbb{E}\left\{\log N_0\right\} - \mathbb{E}\left\{\log N(r,\theta)\right\}. \tag{3.49}$$

By Taylor expansion and assuming $\bar{N}(r,\theta) \gg$,

$$\mathbb{E}\left\{\log N(r,\theta)\right\} \approx \log \mathbb{E}\left\{N(r,\theta)\right\} - \frac{\mathbb{V}\left\{N(r,\theta)\right\}}{2\mathbb{E}\left\{N(r,\theta)\right\}} = \log \bar{N}(r,\theta) - \frac{1}{2\bar{N}(r,\theta)}, \tag{3.50}$$

and therefore by substituting into Eq. (3.49)

$$\mathbb{E}\left\{g(r,\theta)\right\} \approx \log \frac{N_0}{\bar{N}(r,\theta)} = \bar{g}(r,\theta). \tag{3.51}$$

With this, the variance is

$$\mathbb{V}\left\{g(r,\theta)\right\} = \mathbb{E}\left\{\left(g(r,\theta) - \bar{g}(r,\theta)\right)^2\right\} = \mathbb{E}\left\{\left(\log \frac{N(r,\theta)}{\bar{N}(r,\theta)}\right)^2\right\} \overset{\bar{N}\gg}{\approx} \frac{1}{\bar{N}(r,\theta)}. \tag{3.52}$$

Therefore, the variance in a measurement of $\mathscr{R}_2 f$ is inversely proportional to the number of measured photons.

To study the propagation through the reconstruction, we make the following simplifications: We consider a radially symmetric object with constant attenuation coefficient, such that ideally all projections from different angles are equivalent. Further, we only consider the center of the object, i.e. the reconstruction $\hat{f}(0,0)$. Moreover we assume that all measurements $\{g(i_r\Delta_r, \theta_i) : i_r = -N_r, \dots, N_r, \theta_i = \pi(1 - \frac{m}{M}), m = M, \dots, 1\}$ are independent of each other, i.e. there is no systematic error. Then we can write

$$\hat{f}(0,0) \approx \frac{\pi\Delta_r}{M} \sum_{\theta_i} \sum_{k=-K}^{K} g(0,\theta) h(k\Delta_r) \tag{3.53}$$

where h is the spatial implementation of the $|\rho|$ high-pass filter. We further assumed that $g(r,\theta)$ is sufficiently flat around $r = 0$, such that over the support of h simply substitute $g(0,\theta)$. We know that $\mathbb{V}\{g(0,\theta)\} \approx \frac{1}{\bar{N}(0,\theta)}$, and by additivity of variances,

$$\mathbb{V}\left\{\hat{f}(0,0)\right\} \approx \left(\frac{\pi\Delta_r}{M}\right)^2 \frac{M}{\bar{N}(0,\theta)} \sum_{k=-K}^{K} h^2(k\Delta_r). \tag{3.54}$$

With Parseval's Theorem, we can approximate this as

$$\mathbb{V}\left\{\hat{f}(0,0)\right\} \approx \frac{\pi^2\Delta_r}{M\bar{N}(0,\theta)} \int_{-\Omega}^{\Omega} |H(\omega)|^2 \,\mathrm{d}\omega. \tag{3.55}$$

Although many approximations went into Eq. (3.55), it does give valuable insight into what influences the noise level in the reconstruction. We see that the variance is small if the detector spacing Δ_r is small, if we measure a large number M of views, and if the expected number of photons $\bar{N}$ is large. Further, the noise level is proportional to the power spectral density $\int |H|^2$ of h.

3.4.4 Beam Hardening

When discussing the instrumentation in medical *CT*, we saw that the X-Rays are produced by the "continuous" bremsstrahlung and the "discrete" characteristic radiation. A typical spectrum is shown in Fig. 2.6b. Clearly, except for very low energies, the spectrum is spread considerably over all energies, up to the tube voltage. We also mentioned that in practice, the spectrum that "exits" the anode contains has significant power in the low-energy range that is not useful for diagnosis as it would not be able to pass the body at all. Therefore, the spectrum is usually filtered by metal sheets to remove these components. Often this is referred to as *pre-hardening* the spectrum, since the spectrum is shifted towards higher energies, and is therefore *harder*.

Beam Hardening describes the same phenomenon, when it happens in the body that we examine. Recall that the linear attenuation $\mu = \mu(x, E)$ coefficient is a function of the ray energy E, that is, we measure

$$I(r,\theta) = \int_0^\infty S(\epsilon) \exp\left(-\int_0^\infty \mu(r\Theta + s\Theta^\perp, \epsilon)\,\mathrm{d}s\right) \mathrm{d}\epsilon. \tag{3.56}$$

If we let the incident intensity be

$$I_0 = \int_0^\infty S(\epsilon)\,\mathrm{d}\epsilon, \tag{3.57}$$

then, taking into account the energy dependence of μ, the projection integral needs to be modified to

$$F(r,\theta) = -\log\left(\frac{1}{I_0}\int_0^\infty S(\epsilon) \exp\left(-\int_0^\infty \mu(r\Theta + s\Theta^\perp, \epsilon)\,\mathrm{d}s\right) \mathrm{d}\epsilon\right). \tag{3.58}$$

Typically, *soft* (low-energy) X-Rays are more easily absorbed by tissue, such that the X-Rays

are hardened as they pass through the body. The artifacts that arise because of this non-linearity in the measurements are therefore called *beam-hardening artifacts*. The manifestation of beam-hardening artifacts is similar to partial volume artifacts in that, due to the non-linearity, the projections can not properly cancel each other. This gives rise to streak artifacts, which are especially apparent if the X-Rays pass through thick bones.

Note that it is possible to correct for beam-hardening artifacts in a homogeneous theoretical phantom. In fact, since the properties of the linear attenuation coefficient of soft tissue only differ slightly from that of water, scanners are usually calibrated to a generic water phantom to reduce the *cupping effect*. In general however, it is not possible to correct for beam hardening artifacts in unknown objects where the linear attenuation coefficient might vary considerably. On the other hand, the energy dependence of the linear attenuation coefficient is also used advantageously in dual-energy systems.

3.4.5 Scattered Radiation

The *CT* signal model assumes that we can measure a line integral of the linear attenuation coefficient μ. Physically, this means that we assume that photons travel along a straight line, along which they might be absorbed by the medium. In other words, we model the photoelectric effect whilst completely disregarding Compton scattering. It is in fact the case that Compton scattering in general leads to a deterioration of the reconstructed image.

Scattering events change direction (and energy) of the incident photon, thereby deflecting it to be detected by an "off-axis" detector. Clearly, the impact of this scattered radiation is largest in detectors that otherwise would only count very few photons. In fact, scattered radiation can become dominant in regions with strong attenuating structures such as the pelvis [68]. In terms of the effect on the reconstructed image, it is similar to beam hardening artifacts in that areas with high attenuation coefficient are connected by dark streaks.

Scattering artifacts can be reduced by using collimators or anti-scatter grids at the detectors. However, there may be reasons that by construction such collimation is not possible (e.g. fourth-generation scanners), and both traditional collimators and anti-scatter grids in general have other unwanted side-effects [92]. We also want to bring attention to the fact that the number of possible scattering events is obviously proportional to the illuminated volume — that is, in a cone beam setting we expect the influence of scattered radiation to be a lot larger than in a fan-beam setting.

3.4.6 Patient Motion

Motion of the patient with respect to the measurement apparatus is a problem in a range of medical (not necessarily only imaging) procedures. At the same time, even a perfectly compliant patient can not eliminate motion artifacts completely. There are many physiological processes that involve macroscopic movement, which can not be controlled at will. For instance, humans are in general not able to control their heartbeat (and the corresponding pulsating blood flow) or their intestines at will (colon peristalsis). Further, the duration for which patient can hold their breath is in the range of the duration that is needed for some scans. Inconsistent measurement

data over time usually manifests itself as "double images" (i.e. ghost images), or coherent streaks that may extend over the whole image.

Although one may incorporate very simple motion models into the reconstruction problem [80, 88], the only way to eliminate motion artifacts in a general way is to decrease the acquisition time. Of course, since quick acquisition is desired for many reasons, it has drastically decreased over the decades. With modern cone-beam systems, the whole volume can be acquired in the order of seconds, although imaging the beating heart is only possible with Electrocardiography (ECG) gating. However, as discussed before, electron beam *CT* is the fastest method for acquiring slices, which allows to image moving structures in the heart without *ECG* gating.

3.4.7 A Note on 3D

In general, the mechanisms discussed above will lead to very similar artifacts in the two-dimensional and three-dimensional case. The partial-volume effect, scattered radiation, noise and beam hardening will lead to inconsistent measurements, such that streaking artifacts appear in the *FBP* reconstruction. Of course, there exist a whole new class of artifacts that are specific to the three-dimensional reconstruction, especially when considering volume rendering techniques. A well known artifact in volume rendering is the staircasing-effect, where the finite slice thickness leads to a stair-like appearance in regions where there is a strong change along the slice direction. Another artifact that is specific to helical *CT* is scalloping [6], where the intensity of the axial partial volume effect changes with the angle of the measurement. However, this only appears at very high pitch factors and is usually not a problem in clinical practice.

3.4.7.1 Cone Beam Computed Tomography

As already said, three-dimensional *CT* is in general subject to the same artifacts as traditional two-dimensional *CT*. However, the three-dimensional acquisition modalities may reduce the influence of some of the discussed sources of artifacts, while new sources arise. In the case of cone beam *CT*, the possibility to acquire a volume in the order of seconds can almost completely eliminate motion artifacts, but the geometry itself poses a considerable problem for reconstruction.

In the case of "true" three-dimensional reconstruction with real cone beam systems, the problem lies in the fact that with conventional circular source trajectories, it is not possible to acquire the full radon space [12]. Specifically, the three-dimensional Radon space can only be filled completely for object points that intersect the plane that is spanned by the source trajectory. Although three-dimensional analogs to the direct Fourier and *FBP* exist in 3D [24, 37, 58], these methods assume a fully sampled Radon space. A well known three-dimensional reconstruction method that deals with these limitations is the Feldkamp-Davis-Kress (FDK) algorithm [30], which can be adapted for planar and cylindrical detector arrays, and for helical cone beam scanners. In general, cone beam reconstruction, irrespective of the source path, deteriorates with increasing cone aperture.

As already noted, the problem of scattered radiation naturally scales with the illuminated area. An early method for correcting for scatter in cone-beam *CT* was by using "primary modulation" [108], whereby a checkerboard-like source collimator is used such that every other detector in the detector array is shadowed. As the shadowed detectors can only pick up scattered radiation, they can be used to estimate the level of scattered radiation. Coincidentally, this also has the effect of essentially halving the patient dosage. We will discuss this and other strategies of dose reduction in the next section.

3.5 Dose Reduction

From the first clinical *CT* scans in the mid seventies of the last century, up to now *CT* has established itself as one of the most important imaging modalities in clinical practice. With recent advances in electronics, general hardware, and reconstruction techniques in clinical *CT* scanners, the range of applications of *CT* is largely no longer limited by the achievable spatial resolution or patient comfort (i.e. scanning time). Today, the administered patient dose limits the range of applications. In fact, in 2009 it was estimated that medical *CT* accounts for almost half of the ionizing radiation exposure from medical use, which translates to approximately one quarter of the overall average ionizing radiation exposure [71].

By careful assessment of the situation by the physician, it should be guaranteed that the benefits of performing a *CT* scan outweigh the risks associated with it. That is, assuming correct assessment, every *CT* scan performed is a net-benefit to the health care system, and the society that lives along this system. However, that is not to say that no bad consequences exist at all. Although somewhat controversial (see [102] for a very quick review of the potential flaws of the methodology), [10] estimates that radiation exposure from clinical *CT* scans may be directly responsible for 1.5 % to 2 % of cancers in the forthcoming decades.

In any case, it is clear that the cancer risk associated with clinical *CT* is not zero. At this time, with the ability to acquire three-dimensional datasets with sub-millimeter resolution in the order of seconds, it is clear that reducing the dose in *CT* should be one of the top priorities of the *CT* research community. In this section, we will quickly go over the main mechanisms by which we can influence the radiation exposure in clinical *CT*.

3.5.1 Tube Current Reduction

One of the most intuitive ways to control the radiation exposure is to reduce the tube current. It is clear that the patients size and weight influence the dose that is needed to acquire an image of fixed diagnostic value. That is, smaller patients generally require less tube current and therefor less dose to obtain the desired image quality. Historically, the tube current was fixed prior to the measurements by determining the patients weight and looking up the corresponding tube current for a given measurement protocol (e.g. colonography).

In most of today's imaging systems, some form of Automatic Exposure Control (AEC) is implemented. *AEC* aims to control the tube current during acquisition as a function of the detector

signal, such that the image quality stays approximately constant. This may include angular modulation, where the tube current adapts to the potentially changing diameter of the inspected body, as well as longitudinal modulation, where the tube current adapts to the different attenuation at different anatomical regions. For instance, the shoulder as well as pelvis area are known to attenuate the incoming X-Rays stronger than adjacent regions.

Of course, the extent of tube current reduction is limited by the desired Signal-to-Noise Ratio (SNR). As discussed in Section 3.4.3, the detected photons follow a Poisson distribution. In other words, since $\mathbb{V}\{N(r,\theta)\} = \bar{N}(r,\theta)$ the *SNR* "at the detector" (i.e. disregarding the reconstruction) is informally

$$\mathrm{SNR} \propto \frac{\bar{N}}{\sqrt{\bar{N}}} = \sqrt{\bar{N}}. \tag{3.59}$$

As the *SNR* is proportional to the square root of the number of incident photons, it is interesting to consider other means of reducing dose, that may scale better with the number of incident photons.

3.5.2 Angular Undersampling (Few-View CT)

Another well-studied approach for dose reduction is that of undersampling $\mathscr{R}_2 f$ in the angular parameter — that is, by lowering N_θ. Traditionally, with typical analytic reconstruction techniques, this leads to the typical streaking artifacts that significantly inhibit diagnostic value and extent over the whole image domain. However, with the advances in algebraic reconstruction and specifically the field of compressed sensing [26] this has become very popular. A natural extension to angular undersampling in two-dimensional *CT* is to undersample also in the cone-aperture direction in a three-dimensional setting. In [17], the authors proposed a practical multi-slit collimator design, which allows view as well as detector row undersampling.

4 Towards Learning a True Prior

All models are wrong, but some are useful.

George E. P. Box, *Robustness in the Strategy of Scientific Model Building*

Contents

In the previous chapters, we discussed the physical principles of Computed Tomography (CT), went over the image formation, and glimpsed at the differences and advantages or drawbacks of solving the continuous and discretized models. In this chapter, we will focus on reconstruction in the discretized model. Specifically, we will first discuss different strategies for guiding the reconstruction towards plausible solutions given the measurement uncertainties and resulting artifacts that were discussed previously. Then, we will shift our focus towards developing the theory behind the specific approach that we follow throughout this thesis.

Recall that, in essence, we aim to (in some sense) solve the linear system

$$p = Af + v, \tag{4.1}$$

where $p \in \mathscr{P} \subseteq \mathbb{R}^M$ are the projections, $f \in \mathscr{F} \subseteq \mathbb{R}^N$ is the underlying target, and $A : \mathscr{F} \to \mathscr{P}$ is the linear forward operator that describes the acquisition process. We assume that we have knowledge about the distribution of the noise $v \in \mathscr{V} \subseteq \mathbb{R}^M$, or that it can be modeled reasonably well up to some precision. For instance, in clinical *CT* v is typically modeled by

Poisson noise, or may be modeled by spatially heteroskedastic Gaussian noise [94] by utilizing proper pre-processing.

Although the discretized model is stable with respect to perturbations in the input, the reconstruction is in general still strongly influenced by ν and the artifacts discussed in the previous chapter. The following sections will give an overview of different strategies for mitigating the influence of noise and other artifacts. Finally, we will discuss our approach in detail.

4.1 Pre-processing

Considering (4.1), a natural idea is to simply perform denoising on p in the Radon domain. In other words, we want to find

$$\hat{p}_{\text{den}} = \text{den}_p(p) \tag{4.2}$$

where $\text{den}_p : \mathscr{P} \to \mathscr{P}$ is a denoising algorithm and $\hat{p}_{\text{den}}$ estimates the clean, uncorrupted projections. The denoised projections may then be treated without any special considerations. That is, one is completely free to choose any reconstruction algorithm to finally yield the reconstructed image.

This has many advantages, the most obvious of which is that the noise in the projection domain is well understood. This allows to include prior knowledge about the noise distribution in "specialized" approached to the denoising problems. For instance, for (pre-log) Poisson noise, it is natural to use adaptive filters [44, 47]. On the other hand, characterizing the noise in the reconstruction is not as easy, as it strongly depends on the specifics of the reconstruction algorithm. Further, the denoising step is completely independent of the reconstruction following it. That is, the image may be reconstructed with very fast analytical reconstruction algorithms after denoising the projections. Along with this, depending on the specifics, the denoising step itself is usually also very fast with typical approaches. We schematically show pre-processing-based reconstruction approaches in Fig. 4.1.

Throughout the years, many approaches have been proposed, that approximately follow this scheme. This ranges from the aforementioned pre-log Poisson denoising approaches over Total Variation (TV) and wavelet-based approaches [45, 52] to feed-forward Convolutional Neural Networks (CNNs) [34]. Although many of the characteristics of this approach are appealing, the reconstruction quality is typically lacking when compared to that of model-based iterative techniques. Further, it is hard to have a profound intuition of the effect of den_p on the final reconstruction.

4.2 Post-processing

The conceptual opposite of the Radon domain denoising is to perform denoising in the image domain, i.e. after reconstruction. Here, we do not assume access to the projections p, but only to

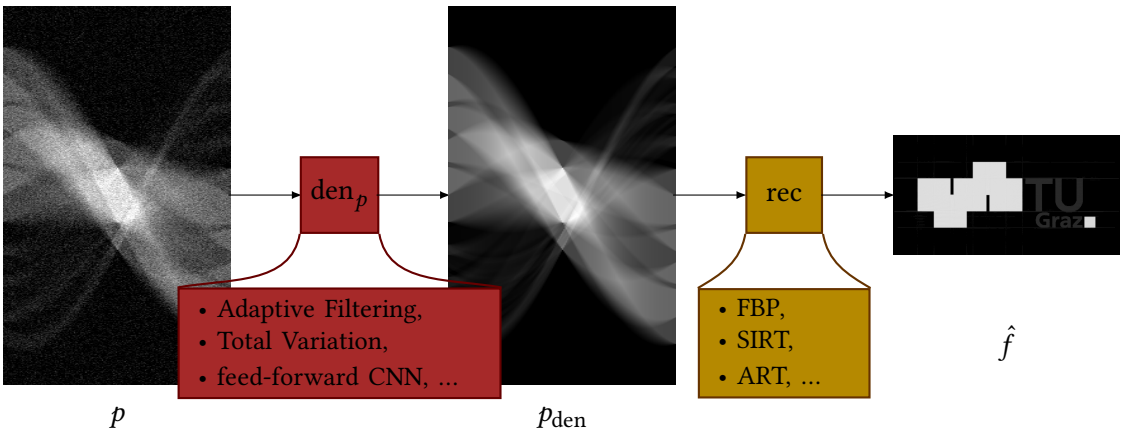


Figure 4.1: Schematic of a pre-processing-based reconstruction pipeline: The denoising algorithm den_p yields the denoised sinogram (middle) from the noisy projections (left), which is subsequently used for reconstruction by any reconstruction algorithm rec to finally yield the reconstructed image (right).

the preliminary reconstruction $\hat{f}$. With this, we simply let

$$\hat{f}_{\text{den}} = \text{den}_f(\hat{f}) \tag{4.3}$$

be the final reconstruction, where $\text{den}_f : \mathscr{F} \to \mathscr{F}$ is an image-domain denoising algorithm. As such, image-domain post-processing can easily extend any existing *CT* workflow, without requiring access to any of the internals. On the other hand, describing the noise in the image-domain is a very challenging task and usually inhibits deriving analytically optimal filtering strategies. Many efforts have been made to adapt typical algorithms to this, e.g. Non-Local Means (NLM) [22, 60, 64] and Block Matching and 3D Filtering (BM3D) [51].

While these traditional methods are good at removing local incoherent noise, low-dose *CT* typically exhibits coherent streaking artifacts. Many learning-based approaches have been proposed to combat such artifacts [19, 20, 61, 101, 106]. Although the results are satisfactory, we emphasize that this discriminative learning setup expects to be applied to images with (at least) very similar corruptions — that is, it assumes a particular forward model as well as reconstruction algorithm. If this condition is not fulfilled, the preliminary reconstructions may exhibit artifacts that the *CNN* can not remove.

4.3 Domain Transform Learning

Another family of reconstruction techniques has emerged very recently with the increase of computational power. Let us denote with $(p, f) \in \mathscr{P} \times \mathscr{F}$ a pair of independent random variables with the associated joint distribution $\mathfrak{D}_D = \mathfrak{D}_p \times \mathfrak{D}_f$ on $\mathscr{P} \times \mathscr{F}$. An intriguing idea is to introduce

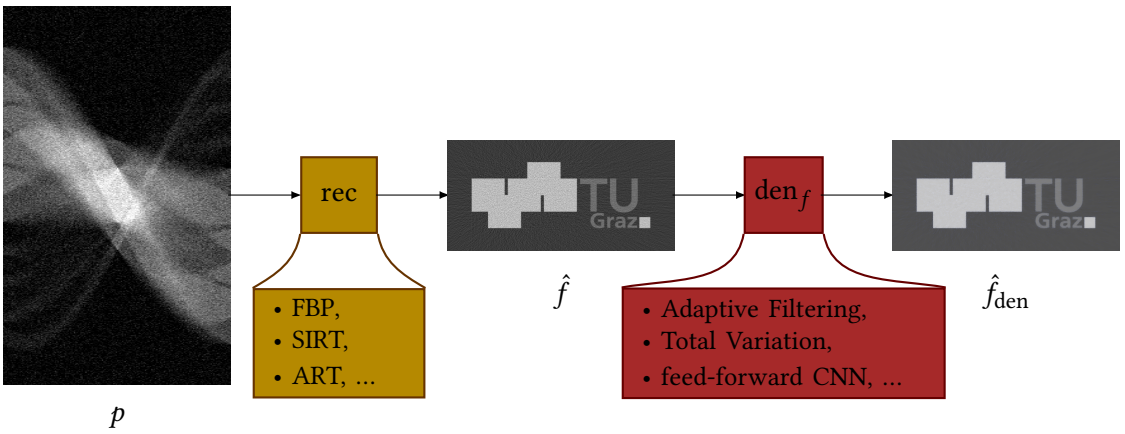


Figure 4.2: Schematic of a post-processing-based reconstruction pipeline: The preliminary reconstruction $\hat{f}$ is acquired by applying some standard reconstruction algorithm to the noisy sinogram p, and subsequently denoised to yield the final reconstruction $\hat{f}_{\text{den}}$.

an appropriately chosen parametric mapping $r : \mathscr{P} \times \Phi \to \mathscr{F}$, such that

$$\hat{f} = r(p, \phi^*), \tag{4.4}$$

where $\phi^* \in \Phi \subseteq \mathbb{R}^P$ is a parameter vector that is learned from data, for instance by minimizing the expected ℓ_2 reconstruction error over $\mathfrak{D}_D$:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathbb{E}_{(p,f)\sim\mathfrak{D}_D}\Big[\|r(p,\phi) - f\|_2^2\Big]. \tag{4.5}$$

Note that in this approach we do not require any domain knowledge at all. That is, we do not require knowledge of the forward operator A or any other specifics of the acquisitions or reconstruction process. We only require a dataset $\mathscr{D}$ that is distributed according to our specific measurement setup. In this sense, this approach is very general and can as stated be applied to any reconstruction task. In [107], the authors applied this framework to a range of medical imaging tasks, including *CT* and Magnetic Resonance Imaging (MRI) using different non-Cartesian sampling patters.

Although the results look promising, there are several problems with this approach. As clearly indicated by Eq. (4.5), this approach is discriminative in nature. As such, one is required to retrain r if the forward model changes, e.g. because a new undersampling pattern is discovered to be advantageous. It also assumes access to both sensor-domain data and image-domain data, and assumes that there exists a known one-to-one association between them. However, especially in the medical domain, data is in general sparse and one can not assume access to a high-quality dataset for every possible scanner geometry and sub-sampling strategy. Further, due to the complete omission of the forward operator A in the reconstruction problem, it is clear that the parametric mapping r has to at least approximate it (more specifically, its "inverse") some-

how. We want to emphasize that the forward models in medical imaging are usually "global" in some sense. For instance, if we consider the Filtered Back-Projection (FBP) example in Fig. 3.2, it is clear that changing one pixel in the sinogram affects the reconstruction along a line over its whole domain. In *MRI*, where the sensor data is acquired in the Fourier domain, changes in the sensor domain change the reconstruction over its whole domain. To account for this fact, the structure of r has to be chosen appropriately. For instance, in the AUTOMAP framework [107], r is a neural network that utilizes fully connected layers in the early stages and convolutional layers in later stages. The idea is to learn an approximate inverse of A in the early stages, and refine the reconstruction by changing the local structure in the later stages. For larger resolutions, this poses a significant computational burden: The number of parameters for one fully connected layer in an *MRI* reconstruction problem of resolution 512×512 is $2 \times 512^4 = 137\,438\,953\,472$, where the factor 2 arises from the complex data. In general it can be concluded that, although appealing in some aspects, reconstruction by means of Eq. (4.4) is not feasible without restrictions on the resolution or specialized architecture considerations.

4.4 Variational Reconstruction

Previously, we outlined some approaches to reconstruction and highlighted some advantages as well as disadvantages. In this section, we will discuss how we aim to solve the reconstruction problem throughout this thesis. First, we will outline the framework of *variational reconstruction*. Then, we will specify our view on the problem and show how we aim to tackle it. In a statistical framework, we will develop an image-domain prior on full *CT*-images, such that we can enjoy all advantages of statistical modeling: Sampling from the prior as well as the posterior, computing different estimators, and rudimentary uncertainty quantification.

Let us again take a closer look at Eq. (4.1). The system is usually overdetermined, since more projections are acquired, than there exist pixels in the reconstruction. Further, the noise v prevents an exact solution in any case. In general we can hope to reconstruct an image f^* that is consistent with the projections in the sense that it minimizes the re-projection error with respect to some metric $d : \mathscr{P} \times \mathscr{P} \to [0, \infty)$. For instance, for the least-squares estimator, we set $d(Af, p) = \|Af - p\|_2^2$ and let

$$f_{\mathrm{LS}}^* = \operatorname*{arg\,min}_{f \in \mathscr{F}} \|Af - p\|_2^2 . \tag{4.6}$$

However, this is in general not satisfactory, as the solution will be guided by the noise v and the reconstruction will strongly depend on the specifics of the forward operator. Therefore, we desire a way to incorporate prior knowledge into the solution, to guide it towards more "physically plausible" solutions.

In inverse problems, *regularization* is the typical way to transform ill-posed problems into related well-posed problems. A well-known regularization technique is due to Tikhonov [97],

who proposed to augment the least-squares objective by

$$f^*_{\mathrm{TK}} = \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{2} \|Af - p\|_2^2 + \frac{\lambda}{2} \|f\|_2^2 \tag{4.7}$$

which hinges on the assumption that $\|f\|_2^2$ is "small" for plausible solutions. Here, $\lambda \in \mathbb{R}^+$ is a parameter that controls the trade-off between f conforming to the measured data p and our prior assumption. However, penalizing the magnitude of the reconstruction is rarely useful in imaging, as the solution would be biased towards low intensity images.

We can generalize Eq. (4.7) to the *variational problem*

$$f^* = \operatorname*{arg\,min}_{f \in \mathscr{F}} \big\{E(f, p) \coloneqq D(f, p) + R(f)\big\}, \tag{4.8}$$

where the energy $E : \mathscr{F} \times \mathscr{P} \to \mathbb{R}$ is composed of a *data-fidelity term* $D : \mathscr{F} \times \mathscr{P} \to \mathbb{R}^+$ which encodes the agreement of the reconstruction with the data, and a *regularizer* $R : \mathscr{F} \to \mathbb{R}$ which penalizes solutions that are far from our prior assumptions on f. We immediately see that Tikhonov regularization sets $D(f, p) = \frac{1}{2} \|Af - p\|_2^2$ and $R(f) = \frac{\lambda}{2} \|f\|_2^2$.

4.4.1 Statistical Interpretation

In order to cast variational methods in the rigorous framework of statistical models, we first state that by Bayes rule, the posterior probability density $\pi_{f|p}(f^*, p)$ of a reconstruction f^* given the data p is

$$\pi_{f|p}(f^*, p) = \frac{\pi_{p|f}(f^*, p)\pi_f(f^*)}{\int_{\mathscr{F}} \pi_{p|f}(\phi, p)\pi_f(\phi)\, \mathrm{d}\phi}, \tag{4.9}$$

where $\pi_{p|f}$ is the *data likelihood* and π_f is the prior. Similar to before, data likelihood describes the agreement between a solution f^* and the measured data p. Assuming an accurate characterization of ν, this is fully determined by the forward model Eq. (4.1). On the other hand, the prior π_f should encode knowledge about the solution itself. Note that the denominator of Eq. (4.9) is intractable for any realistic imaging task. As an example, consider a discrete image of size 4×4, where the pixel values are restricted to be in $\{0, \dots, 255\}$. Then, $|\mathscr{F}| = 4\,294\,967\,296$, which is already in a computationally prohibitive regime. For an image of size 6×6, $|\mathscr{F}| = 4.973\,232 \times 10^{86}$ already surpasses most estimates of the number of baryons (e.g. protons and neutrons) in the observable universe. Luckily, in most applications we do not require to calculate exact probabilities with $\pi_{f|p}$, but we are mostly interested in finding maxima or sampling from it.

Assuming full knowledge of the (maybe unnormalized) posterior $\pi_{f|p}$, we are typically interested in finding the solution which maximizes the posterior. This is known as the popular Maximum-A-Posteriori (MAP) estimator and reads

$$f^*_{\mathrm{MAP}} = \operatorname*{arg\,max}_{f \in \mathscr{F}} \pi_{f|p}(f, p), \tag{4.10}$$

which can be transformed into the negative-log domain as

$$f^*_{\mathrm{MAP}} = \arg\min_{f\in\mathscr{F}} \big\{ -\log \pi_{p|f}(f,p) - \log \pi_f(f) \big\}. \tag{4.11}$$

If we interpret Eq. (4.8) in this context, we see that the data-fidelity term D models the negative data log-likelihood $-\log \pi_{p|f}$, while the regularizer R captures the negative log-prior $-\log \pi_f$.

4.4.2 Hand-crafted Regularizers

While modeling the data-fidelity is usually straight forward, assuming the precise geometry of the scanner is known, finding a good regularizer for *CT* images has been subject to years of research. If we recall Tikhonov regularization, it is now clear that it is not well-suited for imaging applications, since it imposes a "small-magnitude" prior onto f. Therefore, it is interesting to consider other possible choices of R.

A very fruitful idea is that of imposing a "smoothness" prior onto f by penalizing the image gradients. One may initially be tempted to find

$$f^*_{\mathrm{Q}} = \arg\min_{f\in\mathscr{F}} \frac{1}{2}\left\|Af - p\right\|_2^2 + \frac{\lambda}{2}\left\|\mathrm{D}f\right\|_F^2, \tag{4.12}$$

where $\mathrm{D}: \mathbb{R}^N \to \mathbb{R}^{2\times N}$ is a discrete gradient operator and $\|\cdot\|_F$ is the Frobenius norm, which for a matrix $B \in \mathbb{R}^{M\times N}$ is defined as $\|B\|_F = \sqrt{\sum_{m=1}^{M}\sum_{n=1}^{N}(B_{m,n})^2}$. Since $\frac{\lambda}{2}\|\mathrm{D}f\|_F^2$ models the negative log-prior, we can interpret this as assuming that the image gradients of f are normally distributed, with mean 0 and isotropic variance $\frac{1}{\lambda}$. However, this has the effect of smoothing edges in the image, which is generally not desirable. To preserve sharp discontinuities in the image, a very popular approach is to penalize the *TV*. In this case, the regularizer is defined as

$$R(f) = \lambda \left\|\mathrm{D}f\right\|_{2,1} \tag{4.13}$$

where $\|\cdot\|_{2,1}$ denotes the ℓ_1 norm of the ℓ_2 norm with respect to the columns. That is,

$$\left\|\mathrm{D}f\right\|_{2,1} = \sum_{n=1}^{N} \sqrt{((\mathrm{D}f)_{1,n})^2 + ((\mathrm{D}f)_{2,n})^2}. \tag{4.14}$$

The *TV* regularizer and variations of it have been used extensively in the medical imaging domain as well as in problems concerning natural images. For example, in [90], the authors apply the *TV* regularizer to a limited-angle fan-beam reconstruction problem and extended this to a circular cone-beam setup in [91]. A well known drawback of the *TV* regularizer is the staircasing effect [93], and many attempts have been made to overcome this [62, 74, 95, 100, 105]. As an example, we want to highlight the Total Generalized Variation (TGV) [9], which explicitly allows affine image profiles, and has been applied to a sparse-view reconstruction task [74].

4.4.3 Parametric Regularizers

Although the *TV* regularizer is very principled, it still left something to desire in terms of reconstruction quality. The *TGV* regularizer incorporated second-order statistics of the reconstruction and has been shown to improve quality of the reconstruction. Therefore, it is convincing that considering higher-order statistics can improve reconstruction further. However, modeling these statistics by hand becomes increasingly difficult, and it has been advocated quite early that proper modeling of (in this case natural) images should be based on *learning* from data [109].

In this work, we follow the approach of parametrizing our regularizer such that Eq. (4.8) changes to

$$f^* = \arg\min_{f \in \mathcal{F}} \{E(f, p, \phi) := D(f, p) + R(f, \phi)\}, \tag{4.15}$$

where $\phi \in \Phi \subset \mathbb{R}^P$ are the parameters that should be learned from data. Now, the energy $E : \mathcal{F} \times \mathcal{P} \times \Phi \to \mathbb{R}$ measures the compatibility between a reconstruction $f \in \mathcal{F}$ and the data $p \in \mathcal{P}$ given the learned parameters ϕ. In our approach, the data fidelity term remains unchanged, but we parametrize the regularizer $R : \mathcal{F} \times \Phi \to \mathbb{R}$ which encodes prior knowledge on f using the learned parameters ϕ.

Although typically not used in the medical imaging domain, we want to highlight the prolific Fields of Experts (FoE) regularizer due to Roth and Black [84]. For an image $f \in \mathbb{R}^N$, the *FoE* regularizer is defined as

$$R_{\text{FoE}}(f, \phi) = \sum_{n=1}^{N} \sum_{j=1}^{J} \psi((K_j f)_n, w_j) \tag{4.16}$$

where the parameters are summarized as $\phi = (k_j, w_j)_{j=1}^J$, with $k_j \in \mathbb{R}^{a^2}$ a convolution filter of size $a \times a$ corresponding to the linear operators $K_j \in \mathbb{R}^{N \times N}$. In more detail, the potential functions $\psi : \mathbb{R} \times \mathbb{R}^{N_w} \to \mathbb{R}$ are parametrized by the weights $w_j \in \mathbb{R}^{N_w}$, where a typical choice of parametrization may be to use radial basis functions

$$\psi(x, w) = \sum_{k=1}^{N_w} w_k \exp\Big(-\frac{x - \mu_k}{\sigma}\Big). \tag{4.17}$$

Here, μ_k are equidistantly spaced within some interval and σ is chosen a-priori to capture the range of the filter responses. Clearly, the receptive field of the *FoE* regularizer is determined by the filter size a. This is a drawback for medical imaging applications, where artifacts typically appear as coherent streaks over a large footprint.

In this section, we have demonstrated the idea of parametric regularizers and gave as an example the well known *FoE* model. Before we can apply any parametric regularizer to an inference task such as Eq. (4.15), we have to first identify the parameters such that the regularizer encodes useful prior information. In the next sections, we will discuss how we can learn the parameters for a regularizer in an energy of the form Eq. (4.15).

4.5 Parameter Identification

Typical regularizers that are used today in the field of medical imaging or natural image restoration have hundreds, thousands, or hundreds-of-thousands parameters. Clearly, is is not feasible to tune these parameters by hand, and we therefore require ways to learn these parameters from data. In this section, we want to explore ways of learning the parameters of a regularizer for an energy of the form of Eq. (4.15).

First, we want to highlight that in our energy formulation, we only parametrize the regularizer R. R encodes prior information on f, and is independent of any measurements p. As such, one may be tempted to think that we can not incorporate some known one-to-one mappings between p and f for learning ϕ. However, there do exist ways to learn the posterior in a supervised [5, 69] manner, and we want to discuss these first, although this is not the approach that we follow in this thesis.

4.5.1 Bilevel Optimization

By means of bilevel optimization, we can learn the parameters of a regularizer in a supervised manner [59], whereby we solve a higher- and lower-level optimization problem. To be more precise, let $(f_{\mathrm{GT}}, v) \in \mathscr{F} \times \mathbb{R}^M$ denote a pair of independent random variables distributed according to $\mathfrak{D}_D = \mathfrak{D}_f \times \mathfrak{D}_v$. The bilevel optimization approach is to minimize the reconstruction error, where the reconstruction is the minimizer of a lower-level optimization problem, that is

$$\begin{cases} \phi^*_{\mathrm{BO}} = \arg\min_{\phi \in \Phi} \mathbb{E}_{(f_{\mathrm{GT}}, v) \sim \mathfrak{D}_D} \big[\mathscr{L}(f^*(f_{\mathrm{GT}}, v), f_{\mathrm{GT}}) \big], \\ \text{s.t. } f^*(f_{\mathrm{GT}}, v) = \arg\min_{f \in \mathscr{F}} \dfrac{1}{2} \|Af - p\|_2^2 + R(f, \phi). \end{cases} \tag{4.18}$$

Here, $\mathscr{L} : \mathscr{F} \times \mathscr{F} \to \mathbb{R}^+$ is a continuously differentiable loss function and we require R to be twice continuously differentiable. The problem can be solved using Lagrange multiplier theory and implicit differentiation [87]. Note that the lower-level problem typically has to be solved with high precision, which can be computationally expensive.

4.5.2 Truncated Optimization

A significant drawback of bilevel optimization approaches is the computational burden associated with solving the lower level problem. The idea behind truncated optimization [3, 25] is to replace the minimization with an approximate scheme, e.g. fixed-step gradient descent. That is, we aim to solve

$$\begin{cases} \phi^*_{\mathrm{TO}} = \arg\min_{\phi \in \Phi} \mathbb{E}_{(f_{\mathrm{GT}}, v) \sim \mathfrak{D}_D} \big[\mathscr{L}(f^*_{\mathrm{T}}(f_{\mathrm{GT}}, v), f_{\mathrm{GT}}) \big], \\ \text{s.t. } f^*_{\mathrm{t}+1}(f_{\mathrm{GT}}, v) = f^*_{\mathrm{t}}(f_{\mathrm{GT}}, v) - \tau \big(A^*(A f^*_{\mathrm{t}}(f_{\mathrm{GT}}, v) - p) + \nabla_1 R(f^*_{\mathrm{t}}(f_{\mathrm{GT}}, v), \phi) \big). \end{cases} \tag{4.19}$$

for $t = 0, \dots, T-1$ given some initial estimate f_0^*, where A^* denotes the adjoint of A, ∇_N is the gradient w.r.t. the N-th argument, and τ is an appropriately chosen step size. We see that in this fixed-step gradient descent scheme, we can directly propagate the gradient of the loss function through the iterative procedure. It can be shown that under certain conditions, the gradients w.r.t. the parameters of the bilevel problem and the truncated problem converge [66].

At this point, we want to quickly note that while both the bilevel and truncated approach learn the posterior, they can still be interpreted in the framework of energy-based models. An interesting idea is to abandon this framework, and parametrize each of the descent steps in Eq. (4.19) individually. We point the interested reader to [18], where the authors applied this idea to a sparse-view *CT* reconstruction problem.

4.5.3 Maximum Likelihood Learning

The strategies that we discussed in the previous sections aim to, in some sense, learn the posterior distribution directly. While this has some benefits (mainly better discriminative performance), it also has drawbacks: We again assume that we have access to data-image pairs, and the posterior that we learn is tailored to a specific reconstruction problem. In this section, we want to detail a fully generative approach, which does not assume access to measurements at all, and is independent of any particular acquisition setup. We will use this approach to train a full-image regularizer that operates on multiple scales. We detail the specifics of our setup in Section 5.1.

A very well known statistical parameter fitting framework is that of Maximum Likelihood (ML), where we aim to fit our model to the data, such that the likelihood of the data under our model is maximized. Specifically, we interpret *CT* images $f \in \mathscr{F} \subseteq \mathbb{R}^N$ of size $N = N_v \times N_h$ as random variables distributed according to $\mathfrak{D}_f$. To associate a distribution with our regularizer, we follow the maximum-entropy principle [109], such that the probability density reads as

$$\pi_M(f, \theta) = \frac{\exp(-R(f, \phi))}{\int_{\mathscr{F}} \exp(-R(\zeta, \phi)) \, \mathrm{d}\zeta}. \tag{4.20}$$

π_M is often called the *Gibbs-Boltzmann* density of R, and we denote the induced distribution by $\mathfrak{D}_M$.

To find the maximum likelihood estimate $\phi_{\mathrm{ML}}^* \in \Phi$, we can equivalently minimize the expected negative-log likelihood $\mathbb{E}_{f \sim \mathfrak{D}_f}[-\log \pi_M(f, \phi)]$, which amounts to

$$\phi_{\mathrm{ML}}^* = \arg\min_{\phi \in \Phi} \left\{ \Gamma(\phi) \coloneqq \mathbb{E}_{f \sim \mathfrak{D}_f}[R(f, \phi)] + \log\left(\int_{\mathscr{F}} \exp(-R(\zeta, \phi)) \, \mathrm{d}\zeta \right) \right\}. \tag{4.21}$$

The gradient of the *ML* objective with respect to ϕ is found to be

$$\nabla_1 \Gamma(\phi) = \mathbb{E}_{f \sim \mathfrak{D}_f}[\nabla_2 R(f, \phi)] - \int_{\mathscr{F}} \underbrace{\frac{\exp(-R(\zeta, \phi))}{\int_{\mathscr{F}} \exp(-R(\hat{\zeta}, \phi)) \, \mathrm{d}\hat{\zeta}}}_{\pi_M(\zeta, \phi)} \nabla_2 R(\zeta, \phi) \, \mathrm{d}\zeta, \tag{4.22}$$

where we identify the second term to be the expected gradient under the model distribution, such that we arrive at

$$\nabla_1 \Gamma(\phi) = \mathbb{E}_{f^+ \sim \mathfrak{D}_f}\big[\nabla_2 R(f^+, \phi)\big] - \mathbb{E}_{f^- \sim \mathfrak{D}_M}\big[\nabla_2 R(f^-, \phi)\big]. \tag{4.23}$$

Thus, the gradient of the maximum likelihood objective is the difference between the expected gradient of *CT* images and the expected gradient of samples from the model distribution. Therefore, training with Eq. (4.23) has the effect of decreasing the regularization cost of samples from the data distribution, while increasing the regularization cost of "hallucinations" produced by the model. We want to quickly note that this objective also arises in Kullback-Leibler divergence minimization, where the Kullback-Leibler divergence reads

$$(\mathfrak{D}_f \| \mathfrak{D}_M) = \int_{\mathscr{F}} \pi_f(\zeta) \log \frac{\pi_f(\zeta)}{\pi_M(\zeta, \phi)} \, \mathrm{d}\zeta = -\mathrm{H}(\mathfrak{D}_f) - \mathbb{E}_{f \sim \mathfrak{D}_f}[\log \pi_M(f, \phi)]. \tag{4.24}$$

With the entropy H of $\mathfrak{D}_f$ independent of ϕ, we conclude that

$$\arg\min_{\phi \in \Phi} (\mathfrak{D}_f \| \mathfrak{D}_M) = \arg\min_{\phi \in \Phi} \Gamma(\phi). \tag{4.25}$$

We want to emphasize that this method of parameter identification only relies on $\mathfrak{D}_f$. That is, we do not require any measurement data during learning, and our model does not depend on any specific forward operator A or noise distribution $\mathfrak{D}_v$. Since we learn an independent prior (as opposed to the posterior learning from the previous approaches), we also gain the ability to sample our model, such that we can gain valuable insight into what was learned. This comes at the price of an increased computational cost of sampling $\mathfrak{D}_M$, where one typically has to resort to computationally expensive Markov Chain Monte Carlo (MCMC) methods. We will discuss different samplers that could be used in the imaging domain in the next section.

It has been pointed out by Hinton [43] that estimating the expected gradient with respect to the induced model distribution $\mathfrak{D}_M$ is computationally very challenging and often leads to high-variance estimates. Therefore, he proposed to change Eq. (4.23) to

$$\nabla_1 \Gamma(\phi) \approx \mathbb{E}_{f^+ \sim \mathfrak{D}_f}\big[\nabla_2 R(f^+, \phi)\big] - \mathbb{E}_{f^- \sim \mathfrak{D}_{M^T}}\big[\nabla_2 R(f^-, \phi)\big], \tag{4.26}$$

where $\mathfrak{D}_{M^T}$ is the induced model distribution after applying some *MCMC* transition operator (e.g. Gibbs sampling [32]) to $\mathfrak{D}_f$, $T \in \mathbb{N}^+$ times. This approximation, known as the Contrastive Divergence (CD) objective, removes almost all of the computational complexity associated with sampling the model if T is small, and yet yields a reasonable, although biased [13], approximation of the gradient.

CD-based training has historically been used extensively in image restoration problems, with the influential Products of Experts (PoE) [43] and *FoE* [84] models both being trained in this framework. Subsequently, these approaches have fallen out of favor for improved discriminative performance, for instance by using Eq. (4.19). However, recently the *ML* framework has received

increasing attention, especially for its generative capabilities [27, 72]. New models can rival the generative performance of Generative Adversarial Networks (GANs) [36], while preserving the strengths of the probabilistic framework, such as composability, interpretability, and stability due to the lack of an explicit generator network [28, 73].

In this work, we want to build up on the generative capabilities of this approach. That is, we aim to construct a multi-scale network which we train with Eq. (4.26), such that we can sample full-sized images from our model. This is in contrast to, e.g., the *FoE* regularizer, which, although also traditionally trained with Eq. (4.26), can only encode local information.

4.5.3.1 Model Sampling

In the previous section, we derived the objective for fitting a parametric regularizer to data, such that its maximum-entropy distribution maximized the likelihood of the data. The gradient of this loss function with respect to the parameters ϕ of the model, as seen in Eq. (4.23), is composed of two terms: The expected gradient with respect to samples f^+ drawn from the data distribution $\mathfrak{D}_f$, and the negative expected gradient with respect to samples f^- drawn from the induced model distribution $\mathfrak{D}_M$. While the first term is easily approximated given a data set drawn from $\mathfrak{D}_f$, the second term requires sampling from the model, which is achieved by using *MCMC* methods.

There are many approaches to sample from an unnormalized density [11, 5, Chapter 11], which include Gibbs sampling [32], Metropolis-Hastings sampling [42], Hybrid Monte Carlo [29], and Langevin Monte Carlo (LMC) [81]. In this work we restrict ourselves to samplers that 1. make use of the local landscape of the density, and 2. simultaneously update all entries in the underlying random vector. The reasons for this are of practical nature: Samplers that update entries sequentially, such as Gibbs sampling, are impractical for any reasonably sized image (e.g. $512 \times 512 = 262\,144$ entries) in terms of computation. Similarly, using the local landscape (i.e. the gradient) of the density allows for faster convergence of the Markov chains.

We now detail the sampling strategy used in this thesis, which is *LMC* [38, 72, 82, 83]. *LMC* makes use of the gradient of the underlying density during sampling, which improves mixing time when compared to, e.g. Gibbs sampling. Recall that our objective is to sample from $\mathfrak{D}_M$, whose associated density is given by Eq. (4.20), which we repeat as

$$\pi_M(f, \phi) = \frac{\exp(-R(f, \phi))}{\int_{\mathscr{F}} \exp(-R(\zeta, \phi)) \, \mathrm{d}\zeta}.$$

In the general class of Metropolis-Hastings algorithms, there exists a proposal distribution with associated density $\pi_P(f, f^{\rightarrow})$ on $\mathscr{F} \times \mathscr{F}$. A candidate $f^{\rightarrow}$ is drawn from the proposal distribution and the transition is accepted with probability

$$\alpha(f, f^{\rightarrow}) = \begin{cases} \min\left\{\frac{\pi_M(f^{\rightarrow}, \phi)\pi_P(f^{\rightarrow}, f)}{\pi_M(f, \phi)\pi_P(f, f^{\rightarrow})}, 1\right\}, & \text{if } \pi_M(f, \phi)\pi_P(f, f^{\rightarrow}) > 0, \\ 1, & \text{if } \pi_M(f, \phi)\pi_P(f, f^{\rightarrow}) = 0. \end{cases} \tag{4.27}$$

Algorithm 1: *MALA* and *ULA* for sampling from an (unnormalized) density.

Input : Initial point f^0, step size ϵ, Langevin steps T
Output: f^T

for $t = 1, \dots, T$ **do**
  Propose $f^{\rightarrow}$ with Eq. (4.29).
  Compute $\alpha(f^{t-1}, f^{\rightarrow})$ by $\begin{cases} \text{Eq. (4.27)} & \text{for } \textit{MALA}, \\ \alpha(f^{t-1}, f^{\rightarrow}) = 1 & \text{for } \textit{ULA}. \end{cases}$
  Draw $r \sim \mathcal{U}[0, 1]$.
  Set $f^t = \begin{cases} f^{\rightarrow} & \text{if } r < \alpha, \\ f^{t-1} & \text{else.} \end{cases}$
end

This choice of α can be shown to be π_M-invariant, in the sense that

$$\pi_M(f, \phi) = \int_{\mathcal{F}} \pi_M(\xi, \phi) \pi_P(\xi, f) \alpha(\xi, f) \, \mathrm{d}\xi \tag{4.28}$$

and that the transition probabilities converge to π_M. The proposal density we consider is derived from Langevin diffusion: Let f^{t-1} be the current state of the chain, then the proposal distribution takes the form

$$f^t \sim \mathcal{N}(f^{t-1} + \frac{\epsilon}{2} \nabla_1 \log \pi_M(f^{t-1}, \phi), \epsilon \mathrm{Id}_N), \tag{4.29}$$

where $\mathcal{N}(\mu, \Sigma)$ denotes the normal distribution on $\mathbb{R}^N$ with mean μ and covariance matrix Σ.

LMC is also known as the Metropolis adjusted Langevin algorithm (MALA) due to the Metropolis-Hastings step Eq. (4.27). In practice, the Metropolis-Hastings acceptance step is often simply omitted [27, 73], leading to the unadjusted Langevin algorithm (ULA) which only approximately maintains π_M as the invariant distribution. This can be illustrated by the following example taken from [82]: Assume that $\mathfrak{D}_M = \mathcal{N}(0, 1)$ on $\mathbb{R}$ and $\epsilon = 2$, from which it immediately follows that $f^t \sim \mathcal{N}(0, 2)$, $t \in \mathbb{N}^+$. That is, the chain immediately converges, but samples are not distributed according to $\mathfrak{D}_M$. We summarize *MALA* and the unadjusted variant in Algorithm 1.

5

Generative Regularizers for Computed Tomography Reconstruction

> Radiologists who use AI will replace radiologists who don't.
>
> Curtis Langlotz, *RSNA 2017*

Contents

In this chapter, we will detail our model as well as the specifics of the training procedure. Subsequently, we will put our regularizer to test by considering typical reconstruction problems, such as limited-angle and few-view Computed Tomography (CT) reconstruction. Additionally, we will exploit the generative properties of the regularizer. Specifically, we can *sample* the *posterior distribution* for any given problem, and likewise visualize the associated *prior* of our regularizer by drawing samples from it or exploring its modes.

5.1 Model and Training

We start by specifying our model architecture, which is based on [73] and follows a traditional encoder-scheme. The model is schematically shown in in Fig. 5.1. We reduce the spatial dimension of the feature space using strided convolutions, where we blur the kernels following [104]

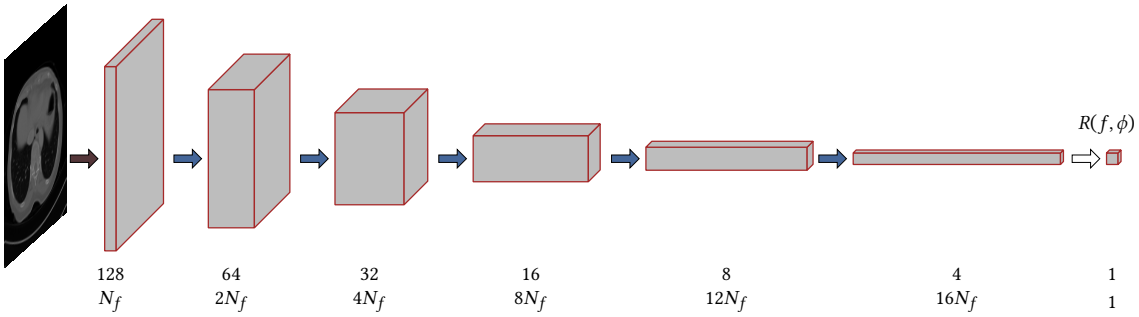


Figure 5.1: Our network follows a typical encoder-structure, where {➡, ➡, ⇨} denote $\{\text{relu} \circ \text{conv}_{3,1}, \text{relu} \circ \text{conv}_{4,2}, \text{conv}_{4,1}\}$, with the subscript specifying the filter size and stride. In the annotations, the upper value shows the spatial resolution of the feature space, and the lower value indicates the number of features.

to avoid aliasing artifacts. All convolutions with the exception of the final, linear layer are followed by the leaky rectified linear unit activation function, with a leak coefficient of 0.05. We use $N_f = 48$ features in the first layer, resulting in in a total of 12 179 905 parameters.

To train our model, we use the Low Dose CT Image and Projection data-set [67], which we subsampled to a spatial resolution of 128×128. We draw samples from our model using the unadjusted Langevin algorithm (ULA) as described in Algorithm 1, which is run for $T = 500$ steps. We further follow the idea of persistent Contrastive Divergence (CD) [96], where we use a replay buffer holding $N_{\text{RB}} = 8000$ past images with a reinitialization chance of 1 %. Upon reinitialization of any given sample in the replay buffer, there is an equal chance of reinitialization with uniform noise or any sample from the training data set. We optimize the Maximum Likelihood (ML) objective Eq. (4.23) using the Adam [54] optimizer with a learning rate of 5×10^{-4}, and set the first and second order momentum variables to $\beta_1 = 0.9$ and $\beta_2 = 0.999$ respectively. To stabilize training, we convolve the data distribution with a Gaussian distribution of standard deviation $\sigma_{\text{data}} = 1.5 \times 10^{-2}$. The training is summarized in Algorithm 2.

For any of the inference tasks we use accelerated proximal gradient descent, as summarized in Algorithm 3. This algorithm makes use of the proximal operator $\text{prox} : \mathscr{F} \to \mathscr{F}$, which for a function $\gamma : \mathscr{F} \to \mathbb{R}$ and $\tau \in \mathbb{R}^+$ is defined as

$$\text{prox}_{\tau\gamma}(\hat{f}) = \arg\min_{f \in \mathscr{F}} \left\{ \tau\gamma(f) + \frac{1}{2} \left\| f - \hat{f} \right\|_2^2 \right\}. \tag{5.1}$$

We point out how we solve the operator for different tasks in their respective sections. Unless stated otherwise, we always run Algorithm 3 with $\alpha = 1 \times 10^{-2}$ and $N_i = 1 \times 10^3$.

5.2 Examining the Regularizer

Before applying our learned regularizer to inference tasks, it is interesting to assess it on a data- and forward model independent basis. The generative *ML* learning scheme allows us to interpret

Algorithm 2: Persistent *CD* training of an energy based model.

Input : data distribution $\mathfrak{D}_f$, data smoothing variance σ^2_{data}, buffer length N_{RB}, reinitialization chance p_{re}, Langevin steps T, initial parameters ϕ, training epochs N_e, Langevin step size ϵ

Output : learned maximum-likelihood parameters ϕ^*

Initialize replay buffer $\mathscr{B} \leftarrow \{u_1, \dots, u_{N_{\text{RB}}}\}$, $u_i \sim \mathscr{U}[0,1]^{128\cdot128}$

for $t = 1, \dots, N_e$ **do**

  Draw $f^+ \sim \mathfrak{D}_f$, $f^0 \sim \mathscr{B}$

  Smooth data samples with $f^+ \leftarrow f^+ + \nu_{\text{data}}, \nu_{\text{data}} \sim \mathscr{N}(0, \sigma_{\text{data}}\text{Id})$

  $\mathscr{B} \leftarrow \mathscr{B} \setminus \{f^0\}$

  Generate f^- with Algorithm 1 using f^0, ϵ, T

  $$\mathscr{B} \leftarrow \mathscr{B} \cup \begin{cases} \{f^-\} & \text{if } r \sim \mathscr{U}[0,1] > p_{\text{re}} \\ \{f^+\} \text{ if } \bar{r} \sim \mathscr{U}[0,1] < 0.5 \text{ else } \{u_i\} \sim \mathscr{U}[0,1]^{128\cdot128} & \text{else} \end{cases}$$

  $\delta_\phi = \nabla_\phi (R(f^+, \phi) - R(f^-, \phi))$

  $\phi \leftarrow \text{Adam}(\delta_\phi)$

end

$\phi^* = \phi$

the learned regularizer directly as a probability density function on the space of images. The probability density is the Gibbs-Boltzmann distribution of R, which reads as

$$\pi_M(f, \theta) = \frac{\exp(-R(f, \phi))}{\int_{\mathscr{F}} \exp(-R(\zeta, \phi)) \, \mathrm{d}\zeta}. \tag{5.2}$$

Naturally, it is interesting to consider the modes of π_M, which by the above equation are easily seen to coincide with the modes of R. We find $f_{\text{mode}} = \arg\min_f R(f, \phi)$ using Algorithm 3 with $D(f, p) = 0$ and $f^0 \sim \mathscr{U}[0,1]^{128\cdot128}$. The proximal operator $\text{prox}_{\alpha D(\cdot, p)}$ reduces to the identity mapping.

We show examples of f_{mode} in Fig. 5.2. To emphasize that these images indeed minimize our learned regularizer, we show example trajectories during minimization along with the corresponding energy in Fig. 5.3. It can be observed that the modes are faithful representations of the training data set. We further want to emphasize that the regularizer is able to learn and retain small details, such as the blood vessels in the lung or the shape of the vertebrae.

The analysis above yields a valuable insight into the regularizer, however it misses a great advantage of the energy-based approach, which is the ability to *sample* the learned prior (or posterior, see Section 5.5). To visualize samples from our learned prior, we run the *ULA* variant of Algorithm 1, where we again choose f^0 as uniform noise. We show the results for $t \in \{1000, 2000, 5000, 20\,000, 39\,900\}$ in Fig. 5.4. The samples exhibit significant change even for very large t. In other words, we are able to traverse modes of the learned prior with the sampling procedure.

In accordance with [73], we find that there is some difference between the modes and samples

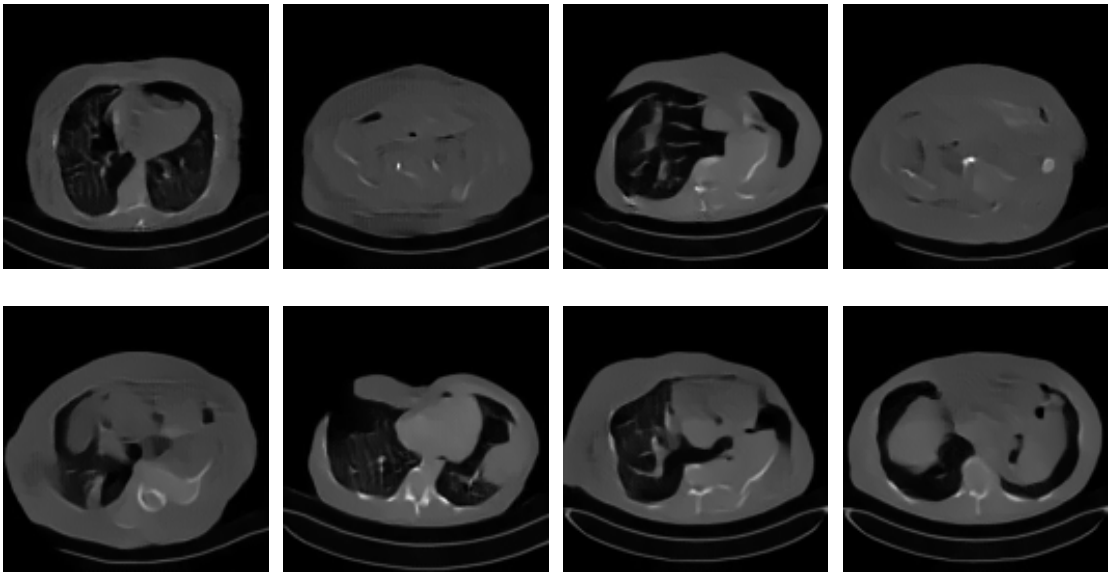


Figure 5.2: Examples of images which locally minimize the learned regularizer R. The images were found using Algorithm 3 with $D(f, p) = 0$ and initializing f^0 with uniform noise. We note that the samples closely resemble the training data, and want to emphasize that the model is able to learn small details in the images, such as blood vessels in the lung.

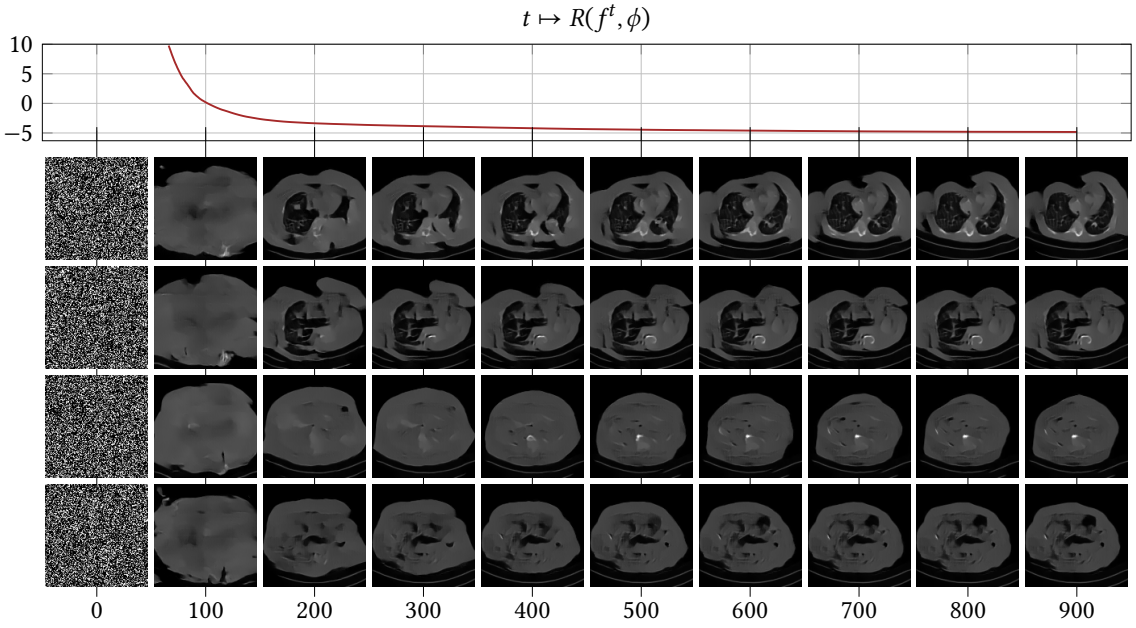


Figure 5.3: Trajectories of the images from uniform noise to $\arg\min_f R(f, \phi)$ along with the corresponding $R(f^t, \phi)$ over the iterations in Algorithm 3.

Algorithm 3: Accelerated Proximal Gradient Descent for minimizing an energy functional.

Input : Initial step size α, data p, initial guess f^0, parameters ϕ, iterations N_i
Output: $f^* = \arg\min_f \{E(f, p, \phi) = D(f, p) + R(f, \phi)\}$

$f^1 = f^0$
for $t = 1, \dots, N_i$ **do**
  $\bar{f} = f^t + \frac{t}{t+3}(f^t - f^{t-1})$
  $g = \nabla_1 R(\bar{f}, \phi)$
  while $R(\text{prox}_{\alpha D(\cdot,p)}(\bar{f} - \alpha g), \phi) > R(\bar{f}, \phi) + g^\top (f^t - \bar{f}) + \frac{1}{2\alpha} \left\| (f^t - \bar{f}) \right\|_2^2$ **do**
    $\alpha \leftarrow \frac{\alpha}{2}$
  end
  $f^{t+1} = \text{prox}_{\alpha D(\cdot,p)}(\bar{f} - \alpha g)$
  $\alpha \leftarrow 2\alpha$
end
$f^* = f^{N_i}$

Algorithm 4: Conjugate Gradient Method to solve $Af = p$.

Input : $A \in \mathbb{R}^{M \times N}$, $f \in \mathbb{R}^N$, $p \in \mathbb{R}^M$, iterations $T \in \mathbb{N}^+$
Output: $f^* = \arg\min_f \|Af - p\|_2^2$

$r^0 = Af - p$
$u = r^0$
for $t = 0, \dots, T - 1$ **do**
  $\alpha = \frac{(r^t)^\top r^t}{u^\top A u}$
  $f \leftarrow f + \alpha u$
  $r^{t+1} = r^t - \alpha A u$
  $u \leftarrow r^{t+1} + \frac{(r^{t+1})^\top r^{t+1}}{(r^t)^\top r^t} u$
end
$f^* = f$

from the learned prior. However, the effect of learning an implicit sampler rather than a faithful density does not seem to be too apparent. That is, the learned modes also largely resemble the training data set, although the samples are a more faithful representation visually.

5.3 Image-Space Inference

In the previous section, we analyzed the learned prior in a data-independent way by minimizing or sampling it. Here, we want to shift our focus to inference tasks in a post-processing-like

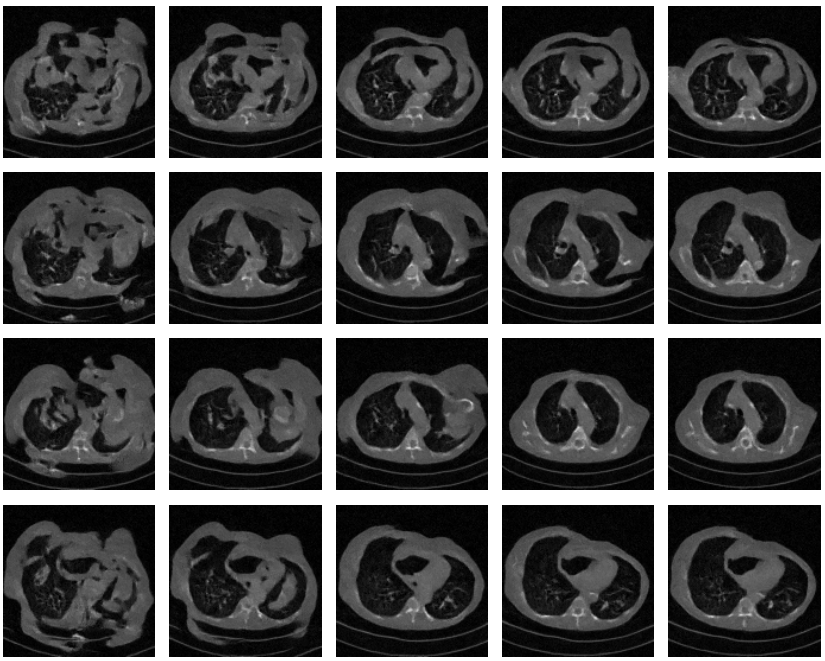


Figure 5.4: Trajectories of the images during Langevin sampling at different time steps. Notice that that high-level features change significantly even for $t \gg$. In other words, our sampling procedure is able to traverse different modes of the learned prior.

framework. Specifically, we consider

$$f^* = \arg\min_{f \in \mathcal{F}} \tilde{D}(f, \tilde{f}) + R(f, \phi) \tag{5.3}$$

where $\tilde{f} \in \mathcal{F}$ is a corrupted *CT* image. To be precise, we let $\tilde{f} = \text{corr}(f_{\text{gt}})$, where corr is an image-space corruption such as Gaussian noise or information loss, and f_{gt} is a fully sampled and uncorrupted *CT* reconstruction. The data fidelity term $\tilde{D} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^+$ is chosen accordingly and does not consider any forward operator A.

For additive white Gaussian noise, we have $\tilde{f} = f_{\text{gt}} + \nu$, $\nu \sim \mathcal{N}(0, \sigma\text{Id})$ and it is known that squared ℓ_2 data term $\tilde{D}(f, \tilde{f}) = \frac{\tilde{\lambda}}{2} \left\| f - \tilde{f} \right\|_2^2$ is optimal. The proximal map $\text{prox}_{\alpha \tilde{D}(\cdot, \tilde{f})}$ can easily be solved in closed form as

$$\text{prox}_{\alpha \tilde{D}(\cdot, \tilde{f})}(f) = \frac{f + \alpha \tilde{\lambda} \tilde{f}}{1 + \alpha \tilde{\lambda}}. \tag{5.4}$$

We compare our method to the Fields of Experts (FoE) [84] and the Total Deep Variation (TDV) [56] models, as well as the Total Variation (TV) regularizer. We find the optimal $\tilde{\lambda}$ using a grid search approach. We optimize the *FoE* model using Algorithm 3, and optimize the *TV* model with a primal-dual algorithm [15]. The *TDV* model is applied in an early stopping framework, as in the original work of [56]. We show quantitative Peak Signal-To-Noise Ratio

Table 5.1: $\mathbb{E}_{f\sim\mathfrak{D}_{\tilde{f}}}[\mathrm{PSNR}(f^*, f)]$ over a test distribution $\mathfrak{D}_{\tilde{f}}$ for denoising.

σ	*TV*	*FoE*	*TDV*	Ours
15	30.71	34.97	37.59	35.93
25	29.00	32.44	34.89	33.39
50	27.78	28.45	30.92	30.04

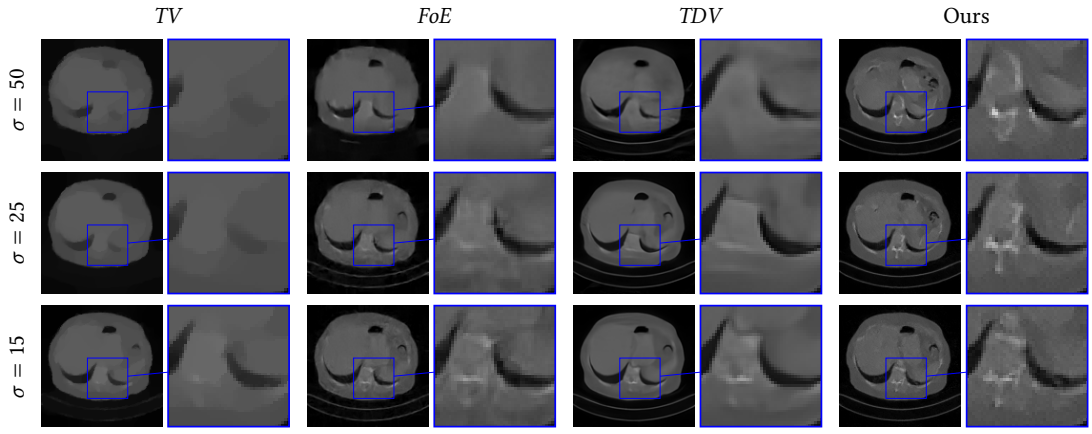


Figure 5.5: Results for a denoising task for $\sigma \in \{15, 25, 50\}$. Our approach retains significantly more detail, however it also hallucinates some structures into the image.

(PSNR) values over an independent test set for $\sigma \in \{15, 25, 50\}$ in Table 5.1, along with visual examples in Fig. 5.5.

Although the *TDV* regularizer beats our approach in the quantitative *PSNR* analysis, we observe that our regularizer is able to retain much more details in the reconstruction. A drawback of our approach is that for high noise levels, where we set $\tilde{\lambda} \ll$ and as such allow the regularizers more freedom, it is able to hallucinate some structures into the image that do not appear in the ground truth reconstruction. However, similar things can be said for the other approaches, which typically lose small structures in the image, due to their implicit or explicit preference for piecewise constant solutions.

As the next task, we consider image inpainting. Here, we assume that we know a subset $\mathcal{I}$ of the image domain Ω exactly, whereas we have no information in $\tilde{\mathcal{I}} = \Omega \setminus \mathcal{I}$. From this, a natural choice for the data term is the Dirac distribution

$$\tilde{D}(f, \tilde{f}) = \delta_{\mathcal{I}}(f, \tilde{f}) = \begin{cases} \infty & \text{if } f \neq \tilde{f} \text{ anywhere in } \mathcal{I}, \\ 0 & \text{else,} \end{cases} \tag{5.5}$$

Table 5.2: $\mathbb{E}_{f\sim\mathfrak{D}_{\tilde{f}}}[\mathrm{PSNR}(f^*, f)]$ over a test distribution $\mathfrak{D}_{\tilde{f}}$ for different inpainting tasks.

Task	p_i	*TV*	*FoE*	*TDV*	Ours
Line	0.5	30.06	31.45	11.68	31.62
	0.8	23.58	24.20	9.82	23.70
	0.9	18.28	19.88	9.40	20.55
Pixel	0.5	34.71	41.46	20.91	40.03
	0.8	29.59	33.02	15.22	33.44
	0.9	27.43	29.47	12.10	29.42

where the proximal operator decays to the projection

$$\mathrm{prox}_{\alpha\delta_{\mathcal{I}}(\cdot,\tilde{f})}(f) = \begin{cases} \tilde{f} & \text{in } \mathcal{I}, \\ f & \text{else.} \end{cases} \tag{5.6}$$

Specifically, we will consider two cases that differ in the nature of $\mathcal{I}$:

1. Line inpainting: For each horizontal line in Ω, there is a chance p_i that its pixels are in $\tilde{\mathcal{I}}$.
2. Pixel inpainting: Each pixel in Ω has a chance p_i to be in $\mathcal{I}$.

We show the *PSNR* values over an independent test set for $p_i \in \{0.5, 0.8, 0.9\}$ in Table 5.2, and show qualitative results in Fig. 5.6. We want to bring attention to the fact that *PSNR* analysis becomes more and more meaningless as p_i increases, since there is more and more ambiguity and natural reconstructions may be far from the initial image. We find that our approach leads to the most natural and detailed reconstructions, especially as the percentage of missing information increases, that is as $p_i \to 1$. We attribute this to the fact that our regularizer has a global receptive field and as such can draw correspondences over large distances. Further, we see that the *TDV* regularizer, which was trained discriminatively on an additive Gaussian denoising tasks fails catastrophically at this task, as it tries to smooth image and hallucinates very unnatural structures into the image.

5.4 Model-Based Reconstruction

In this section, we will apply our learned prior to *CT* reconstruction tasks. Specifically, we will focus on limited-angle and few-view *CT* reconstruction. Recall that our forward model is given by

$$p = Af + \nu, \tag{5.7}$$

and we are trying to find

$$f^* = \operatorname*{arg\,min}_{f\in\mathcal{F}} \{E(f, p, \phi) \coloneqq D(f, p) + R(f, \phi)\}. \tag{5.8}$$

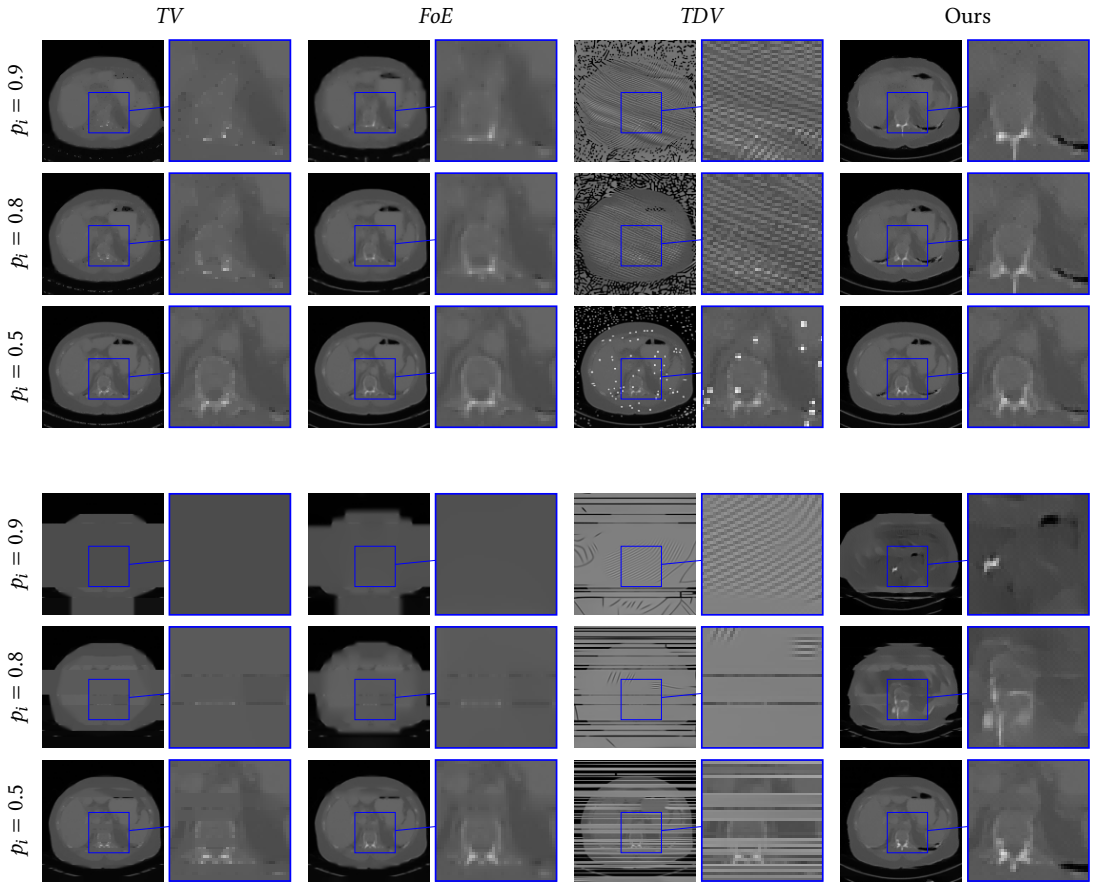


Figure 5.6: Results for an inpainting task for $p_i \in \{0.5, 0.8, 0.9\}$. The top three rows show a pixel-wise inpainting problem, and the bottom three rows depict a line-wise inpainting problem. We find that our approach leads to the most natural and detailed reconstructions visually.

In what follows, we use the ASTRA toolbox [1, 2] to discretize our forward operator A using the scheme detailed in Fig. 3.6. We also use their implementation of the traditional reconstruction algorithms, namely Simultaneous Algebraic Reconstruction Technique (SART) and Filtered Back-Projection (FBP). For all experiments, we let ν be 0.1 % Gaussian noise and use

$$D(f, p) = \frac{\lambda}{2} \|Af - p\|_2^2, \tag{5.9}$$

where we find λ using grid-search. We run the *SART* algorithm for 5000 iterations, and solve the *TV* problem using 400 iterations of the primal-dual algorithm [15]. We solve the proximal maps with Algorithm 4 using $T = 10$ iterations.

Recall that, for Algorithm 3, we need to compute

$$\operatorname{prox}_{\alpha D(\cdot, p)}(\hat{f}) = \arg\min_{f \in \mathcal{F}} \left\{ \rho(f) := \frac{\alpha\lambda}{2} \|Af - p\|_2^2 + \frac{1}{2} \left\| f - \hat{f} \right\|_2^2 \right\}. \tag{5.10}$$

Table 5.3: $\mathbb{E}_{f\sim\mathfrak{D}_{\bar{f}}}[\mathrm{PSNR}(f^*, f)]$ over a test distribution $\mathfrak{D}_{\bar{f}}$ for limited-angle ($\theta \in [0, \frac{\pi}{2}]$) reconstruction.

Method	*FBP*	*SART*	*TV*	Ours
$\mathbb{E}_{f\sim\mathfrak{D}_{\bar{f}}}[\mathrm{PSNR}(f^*, f)]$	19.05	27.72	29.67	34.21

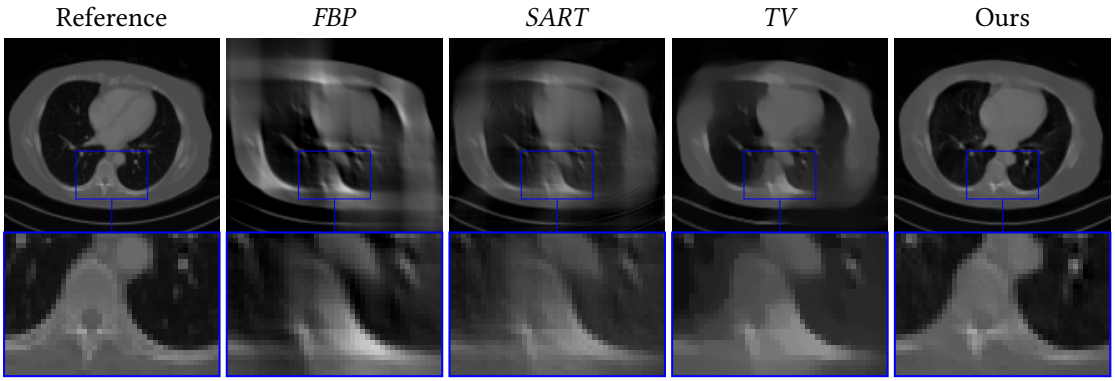


Figure 5.7: Comparison between *FBP*, *SART*, *TV*, and our method for limited-angle ($\theta \in [0, \frac{\pi}{2}]$) *CT* reconstruction. Our model is able to faithfully reconstruct the image, whereas the other methods are not able to fully remove the limited-angle smearing artifacts.

By construction, we know that $\arg\min_f \rho(f) \in \mathcal{F}$, such that by convexity of ρ

$$f^*_{\mathrm{prox}} = \mathrm{prox}_{\alpha D(\cdot,p)}(\hat{f}) \Leftrightarrow \nabla_1 \rho(f^*_{\mathrm{prox}}) = 0. \tag{5.11}$$

The condition Eq. (5.11) can be rewritten as

$$(\alpha\lambda\bar{A}A + \mathrm{Id}_N) f^*_{\mathrm{prox}} = \alpha\lambda\bar{A}p + \hat{f}, \tag{5.12}$$

where $\bar{A}$ is the adjoint of A. This can be solved using Conjugate Gradient (CG) as described in Algorithm 4.

We first consider the problem of limited angle *CT* reconstruction. Here, the sinogram is acquired only over a range of angles, which typically is considerably smaller than π rad. Specifically, we sample 270 projections uniformly spaced over $[0, \frac{\pi}{2}]$, with 362 detectors using a detector spacing of 1 pixel. With our images discretized on a 128×128 pixel grid, this results in a forward operator $A \in \mathbb{R}^{(362\cdot270)\times(168\cdot168)}$. We compare our method to *FBP*, *SART* and *TV* reconstruction quantitatively in Table 5.3 and show some examples of the reconstruction in Fig. 5.7.

Our model is able to satisfactorily reconstruct the image, whilst the traditional unregularized reconstruction algorithms fail to faithfully restore the image. This is especially apparent at contours where the projection streaks are not cancelled by other projections, such as in the upper left or lower right corners in Fig. 5.7. Although *TV* regularization helps alleviate this somewhat, the structures still look unnatural and it is not able to fully restore the contours. On the other

Table 5.4: $\mathbb{E}_{f\sim\mathfrak{D}_{\bar{f}}}[\mathrm{PSNR}(f^*, f)]$ over a test distribution $\mathfrak{D}_{\bar{f}}$ for few-view *CT* reconstruction using $N_\theta \in \{20, 30, 50, 100\}$.

N_θ	*FBP*	*SART*	*TV*	Ours
100	37.15	43.86	46.77	49.47
50	33.12	37.05	40.21	45.06
30	28.78	33.04	35.33	41.65
20	25.24	30.55	31.77	38.48

hand, the global receptive field of our approach allows to recover an image that is consistent with the training data. Further, we want to point out that our model does not hallucinate any unwanted structures into the reconstruction, as the data term only allows changes to the image that are consistent with the measurement data.

An interesting avenue for dose reduction, which has come to light with the advent of Compressed Sensing (CS)-based methods [26], is to perform angular undersampling [90, 91]. That is, as discussed in Section 3.4, instead of densely sampling the Radon space in the rotational parameter we only acquire a few views, typically in the range of 100. For our experiments, we now sample N_θ views uniformly spaced over $[0, \pi]$, with 362 detectors each using a detector spacing of 1 pixel. The expected *PSNR* values of our method along with traditional reference methods over a test set for $N_\theta \in \{20, 30, 50, 100\}$ are shown in Table 5.4. We further show a visual comparison in Fig. 5.8.

We again observe that our model is able to reconstruct the image satisfactorily, with small details such as the blood vessels in the lung visible even for N_θ as low as 20. For such strong undersampling, we observe that *FBP* reconstruction introduces very strong streaking artifacts, which are less apparent in the *SART* reconstruction at the cost of an oversmoothed reconstruction. *TV* regularization is able to restore sharp edges from this, however we find that recovering a sharp images comes at the cost of losing almost all detail in the image, with only few gray levels remaining. Our approach is able to eliminate the streaking artifacts whilst also retaining a good level of detail in the reconstruction. The quantitative analysis in Table 5.4 also shows that our model outperforms the other reconstruction methods by a large margin.

5.5 Posterior Sampling

In the previous sections, we applied our learned prior to typical *CT* reconstruction tasks. There, we treated our regularizer as a point estimator, where the reconstruction is given as the Maximum-A-Posteriori (MAP) of the Gibbs-Boltzmann distribution of the energy. In other words, we computed

$$f^*_{\mathrm{MAP}} = \operatorname*{arg\,min}_{f\in\mathcal{F}} \left\{E(f, p, \phi) \coloneqq D(f, p) + R(f, \phi)\right\}. \tag{5.13}$$

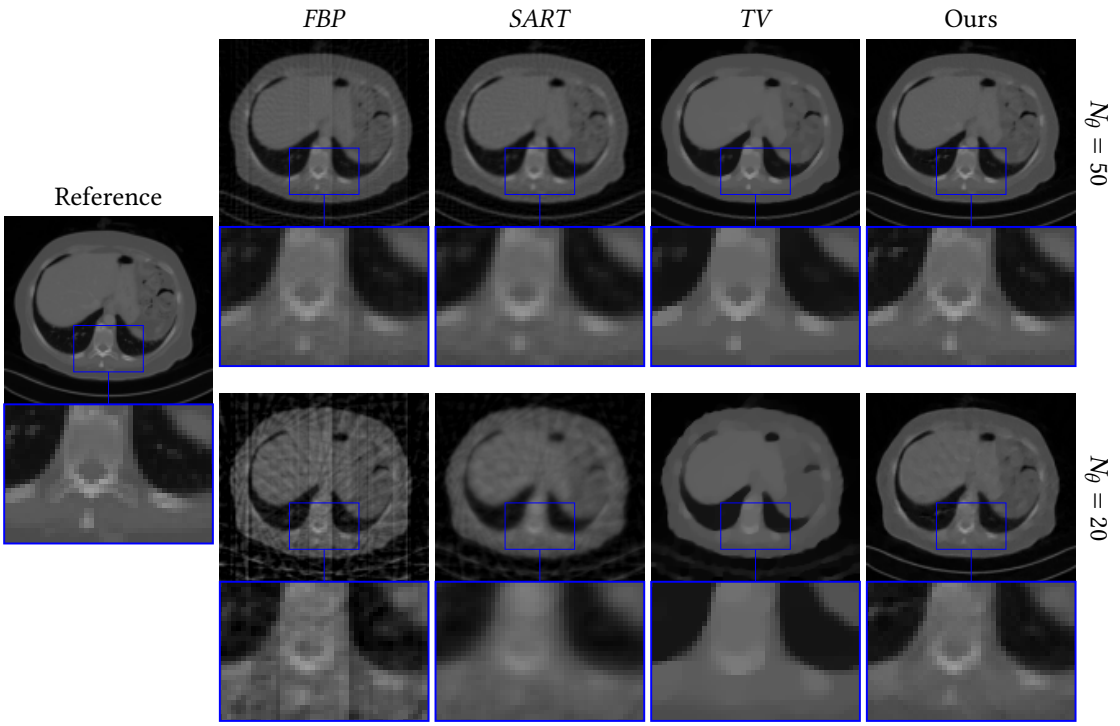


Figure 5.8: Comparison between *FBP*, *SART*, *TV*, and our method for few-view *CT* reconstruction. Our model is able to completely remove the streaking artifacts while retaining a satisfactory level of detail.

However, our formulation allows us to leverage the full posterior distribution. That is, we may also consider other estimators for the optimal solution, such as the expectation over the posterior

$$f_{\mathrm{E}}^{*} = \mathbb{E}_{f\sim\mathfrak{D}_E}[f] = \int_{\mathscr{F}} f \frac{\exp(-E(f,p,\phi))}{\int_{\mathscr{F}} \exp(-E(\bar{f},p,\phi))\,\mathrm{d}\bar{f}}\,\mathrm{d}f, \tag{5.14}$$

where $\mathfrak{D}_E$ is the Gibbs-Boltzmann distribution of $E(f,p,\phi)$ given the parameters ϕ. Clearly, it is intractable to solve Eq. (5.14) analytically, but we may approximate it using Langevin Monte Carlo (LMC), or simply visually examine samples from the posterior.

To visualize this, we use the same few-view and limited-angle forward operator as in the previous section using $N_\theta = 30$ and $\theta \in [0, \frac{\pi}{2}]$ respectively. We inject noise into Algorithm 3 after computing the proximal map in Line 8, where we draw the noise from $\mathcal{N}(0, \alpha\beta\mathrm{Id})$, where β is chosen as $\beta = 2 \times 10^{-3}$. We show a sampling trajectory along with the mean and variance over all iterations in Fig. 5.9. We observe high variance around small structures such as the vertebrae and the blood vessels in the lung, and along contours for few-view reconstruction as well as regions of high ambiguity in the limited-angle reconstruction (c.f. Fig. 5.7).

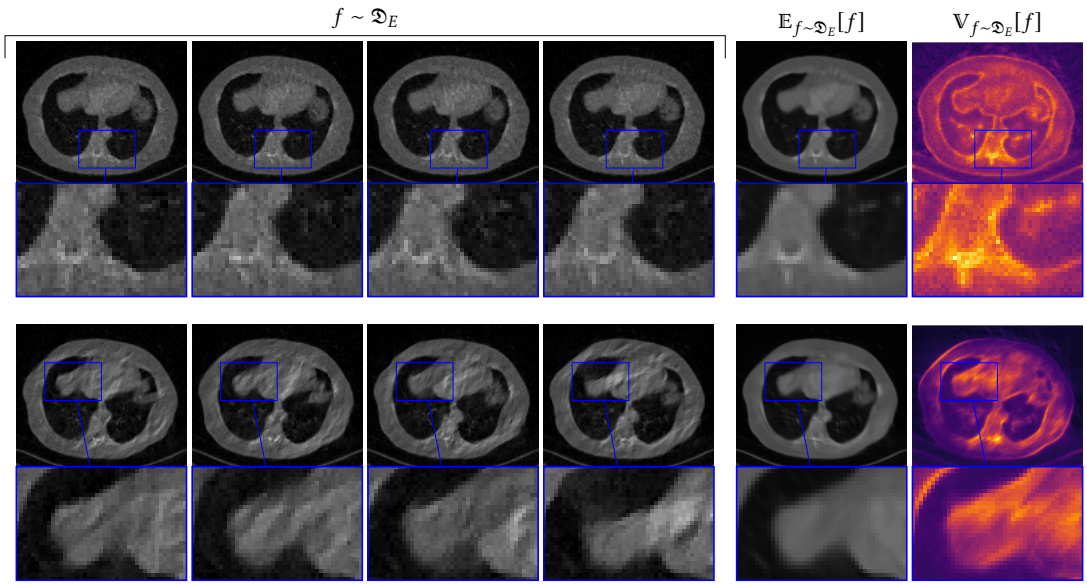


Figure 5.9: Sampling the posterior of a few-view ($N_\theta = 30$, top) and limited-angle ($\theta \in [0, \frac{\pi}{2}]$, bottom) *CT* reconstruction problem: The four images on the left show different samples during the sampling process, the two images on the right show the expected sample and variance of the posterior distribution respectively.

5.6 A Note on Scale-Non-Invariance and Out-Of-Distribution Application

In the architecture of our regularizer, we explicitly consider the image size, as the repeated strided convolutions directly map from 128×128 to a scalar. This means that our network is explicitly not invariant with respect to the scale of the image, or in other words, we have learned a prior on *CT* images of exactly this resolution. However, since the network is composed only of convolutional layers (as opposed to fully connected layers), we may apply it to different resolution images.

To examine the behavior in such cases, we consider an image-space denoising task for different resolutions 256×256 and 512×512. We solve the denoising problem using accelerated proximal gradient descent for $\sigma \in \{25, 50\}$, and show the results along with a sample of our model on these resolutions in Fig. 5.10. We see that, even for a denoising task with $\sigma = 25$, the regularizer is not able to restore the image satisfactorily, as it hallucinates new structures into the image while the noise is is still apparent. In general, due to the design of our regularizer, we can not expect it to work on resolutions other than the training resolution.

To study the behavior on out-of-distribution samples, we perform the following denoising experiment: We let

$$f_\kappa = \mathrm{rot}_\kappa(f) + \nu, \tag{5.15}$$

where $\mathrm{rot}_\kappa : \mathscr{F} \to \mathscr{F}$ is the bilinear rotation operator of angle κ and $\nu \sim \mathcal{N}(0, \sigma\mathrm{Id})$, $\sigma = 25$. We determine the optimal $\tilde{\lambda}$ using manual grid search on f_0, and show expected *PSNR* values

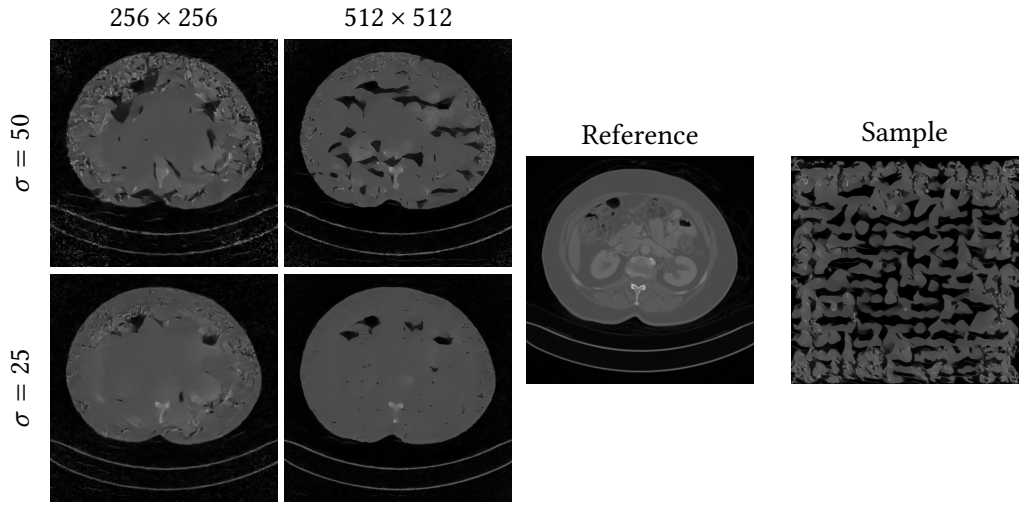


Figure 5.10: Experiments on different scales: Results for Gaussian denoising and a sample of our prior on 512×512. Since our regularizer is not scale-invariant, the results are not satisfactory.

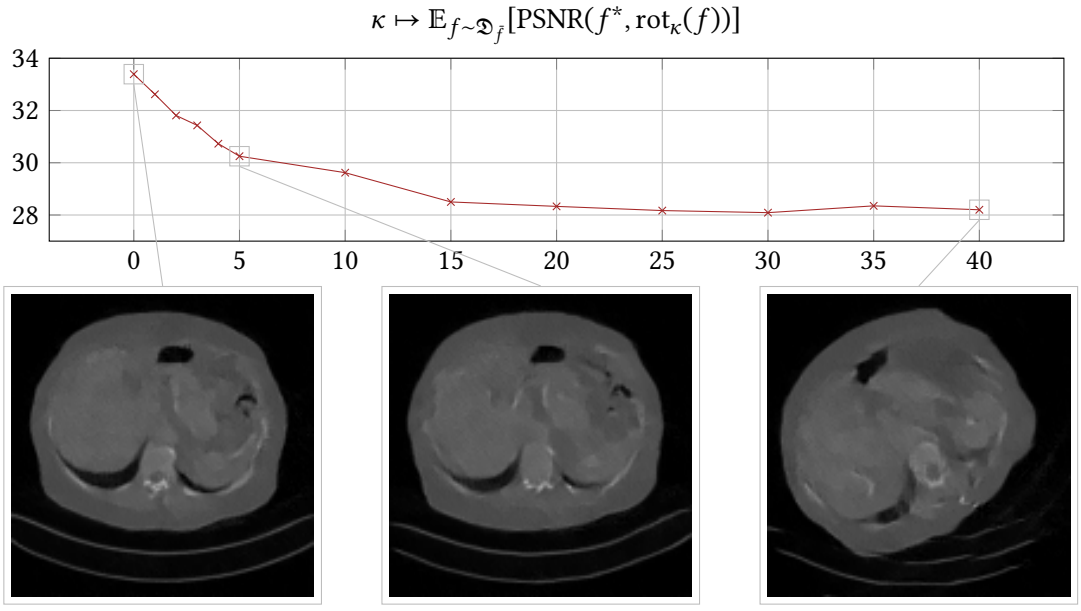


Figure 5.11: Performance of the regularizer on out-of-distribution data: For denoising rotated images, the *PSNR* quickly decays even for small rotations.

in Fig. 5.11 for $\kappa \in \{1°, 2°, 3°, 4°, 5°, 10°, \dots, 40°\}$. We can see that the *PSNR* value quickly decreases even for small κ. For larger κ, the model introduces unnatural structures in the image visually, which goes hand in hand with a strong drop in quantitative performance. In general, since our model is a strong prior only for "upright" *CT* images, performance quickly degrades even for small affine transformations on the images drawn from the training distribution.

6 Conclusion and Outlook

A theory that explains everything, explains nothing.

Karl Popper

Contents

6.1 Conclusion

In this work, we introduced a novel fully generative approach to learn a regularizer for *CT* reconstruction. Our energy-based formulation allows us to apply this regularizer to a multitude of reconstruction task, such as limited-angle or few-view *CT*, as well as image-space restoration tasks such as denoising. The learned regularizer is able to outperform traditional reconstruction algorithms in all tasks, sometimes by a large margin. The advantages of the global receptive field of our regularizer along with the generative training are especially apparent in limited-angle and few-view reconstruction tasks. There, the regularizer is able to find solutions that are consistent with the measured data and exhibit the global structure of the fully-sampled references.

Further, we can cast the energy-based model into a statistical framework, which allows us to leverage the rich theory of statistical models. Specifically, we can visualize our prior by means of computing the modes of its Gibbs-Boltzmann distribution, or by drawing samples from it. This gives valuable insight into what the regularizer has learned, which is critical in the medical domain where interpretability is exceptionally important. On the same note, for any reconstruction task, we may not only compute one point estimate by means of the *MAP* solution, but we can draw samples from the posterior. This allows us to also compute the expectation as well as the variance of the posterior, which can be interpreted as a rudimentary uncertainty quantification.

6.2 Outlook

The probabilistic interpretation opens an avenue for domain experts to gain insight into what the regularizer prefers, and if this is helpful at all in the reconstruction task. However, there are some limitations to this approach: Due to the computational burden of model sampling during training, such approaches have traditionally not been used. With recent advances in computational power and efficiency, it is now possible to train models on resolutions close to clinical practice.

The *ML* training however is not only computationally expensive, but also suffers from instability during training. An interesting question is if the training can be stabilized by injecting "discriminative" knowledge. For instance, one may be tempted to train a regularizer that is a capable generative model, while it is simultaneously able to classify the z-axis position of the slice. Similarly, it may be interesting to train a network with *ML* as well as a discriminative loss, that is obtained by a segmentation task. This simultaneous hybrid training of one model has recently been shown to boost discriminative performance, and it may be a way to stabilize the *ML* training. Stable training would allow us to train larger, more expressive networks, that could be applied to larger resolutions.

Although our regularizer is a very strong global prior for *CT* images of a certain size, the experiments on other scales and out-of-distribution data showed that the performance quickly degrades for images with other high-level characteristics. Therefor, a possible avenue of future work is to train a regularizer in a scale-invariant manner, for instance by treating the scale as a latent variable. Further, our model only contains convolutional layers, which may not be optimal for such tasks. In light of the global receptive field of our model, it may be interesting to include attention layers, that are inherently global. We believe that by combining hybrid training with recent advances in the neural network community, such as attention layers, it is possible to train highly expressive models. In general, we feel that learning fully generative models to be used for regularization is a very interesting topic for future research.

A

List of Acronyms

AEC	Automatic Exposure Control
ART	Algebraic Reconstruction Technique
BICAV	Block Iterative Component Averaging
BM3D	Block Matching and 3D Filtering
CD	Contrastive Divergence
CG	Conjugate Gradient
CNN	Convolutional Neural Network
CS	Compressed Sensing
CT	Computed Tomography
ECG	Electrocardiography
FBP	Filtered Back-Projection
FDK	Feldkamp-Davis-Kress
FoE	Fields of Experts
GAN	Generative Adversarial Network
LMC	Langevin Monte Carlo
MALA	Metropolis adjusted Langevin algorithm
MAP	Maximum-A-Posteriori
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MRI	Magnetic Resonance Imaging
NLM	Non-Local Means
OS-SQS	Ordered Subset Separable Quadratic Surrogates
PoE	Products of Experts
PSNR	Peak Signal-To-Noise Ratio
SART	Simultaneous Algebraic Reconstruction Technique
SBP	Simple Back-Projection
SIRT	Simultaneous Iterative Reconstruction Technique

SNR	Signal-to-Noise Ratio
TDV	Total Deep Variation
TGV	Total Generalized Variation
TV	Total Variation
ULA	unadjusted Langevin algorithm

Bibliography

[1] W. van Aarle et al. "Fast and flexible X-ray tomography using the ASTRA toolbox." In: *Optics Express* 24.22 (Oct. 2016), p. 25129. DOI: 10.1364/oe.24.025129 (cited on page 65).

[2] W. van Aarle et al. "The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography." In: *Ultramicroscopy* 157 (Oct. 2015), pp. 35–47. DOI: 10.1016/j.ultramic.2015.05.002 (cited on page 65).

[3] A. Barbu. "Training an Active Random Field for Real-Time Image Denoising." In: *IEEE Transactions on Image Processing* 18.11 (Nov. 2009), pp. 2451–2462. ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2009.2028254 (cited on page 51).

[4] J. F. Barrett and N. Keat. "Artifacts in CT: Recognition and Avoidance." In: *RadioGraphics* 24.6 (Nov. 2004), pp. 1679–1691. DOI: 10.1148/rg.246045065 (cited on page 2).

[5] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738 (cited on pages 51, 54).

[6] C. A. Blanck. *Understanding Helical Scanning*. eng. 1998. ISBN: 9780683303049 (cited on page 40).

[7] K. Borozdin et al. "Cosmic Ray Radiography of the Damaged Cores of the Fukushima Reactors." In: *Physical Review Letters* 109 (15 Oct. 2012), p. 152501. DOI: 10.1103/PhysRevLett.109.152501 (cited on page 6).

[8] R. N. Bracewell. *The Fourier Transform and Its Applications*. McGraw-Hill Series in Electrical Engineering. McGraw-Hill, 1986. ISBN: 9780070070165 (cited on page 23).

[9] K. Bredies, K. Kunisch, and T. Pock. "Total Generalized Variation." In: *SIAM Journal on Imaging Sciences* 3.3 (Jan. 2010), pp. 492–526. DOI: 10.1137/090769521 (cited on pages 1, 49).

[10] D. J. Brenner and E. J. Hall. "Computed Tomography — An Increasing Source of Radiation Exposure." In: *New England Journal of Medicine* 357.22 (Nov. 2007), pp. 2277–2284. DOI: 10.1056/nejmra072149 (cited on page 41).

[11] S. Brooks et al. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011. ISBN: 9781420079418 (cited on page 54).

[12] T. M. Buzug. *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*. eng. 2008. ISBN: 9783540394075 (cited on pages 1, 6, 25, 40).

[13] M. Á. Carreira-Perpiñán and G. Hinton. "On Contrastive Divergence Learning." In: *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*. Ed. by R. G. Cowell and Z. Ghahramani. Vol. R5. Proceedings of Machine Learning Research. PMLR, June 2005, pp. 33–40 (cited on page 53).

[14] Y. Censor, D. Gordon, and R. Gordon. "BICAV: a block-iterative parallel algorithm for sparse systems with pixel-related weighting." In: *IEEE Transactions on Medical Imaging* 20.10 (2001), pp. 1050–1060. DOI: 10.1109/42.959302 (cited on page 33).

[15] A. Chambolle and T. Pock. "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging." In: *Journal of Mathematical Imaging and Vision* 40.1 (Dec. 2010), pp. 120–145. DOI: 10.1007/s10851-010-0251-1 (cited on pages 62, 65).

[16] R. H. Chan, M. Tao, and X. Yuan. "Constrained Total Variation Deblurring Models and Fast Algorithms Based on Alternating Direction Method of Multipliers." In: *SIAM Journal on Imaging Sciences* 6.1 (Jan. 2013), pp. 680–697. DOI: 10.1137/110860185 (cited on page 1).

[17] B. Chen et al. "SparseCT: System Concept and Design of Multislit Collimators." In: *Medical Physics* 46.6 (May 2019), pp. 2589–2599. DOI: 10.1002/mp.13544 (cited on pages 1, 42).

[18] H. Chen et al. "LEARN: Learned Experts' Assessment-Based Reconstruction Network for Sparse-Data CT." In: *IEEE Transactions on Medical Imaging* 37.6 (2018), pp. 1333–1347. DOI: 10.1109/TMI.2018.2805692 (cited on page 52).

[19] H. Chen et al. "Low-dose CT via Convolutional Neural Network." In: *Biomedical Optics Express* 8.2 (Jan. 2017), p. 679. DOI: 10.1364/boe.8.000679 (cited on page 45).

[20] H. Chen et al. "Low-Dose CT With a Residual Encoder-Decoder Convolutional Neural Network." In: *IEEE Transactions on Medical Imaging* 36.12 (2017), pp. 2524–2535. DOI: 10.1109/TMI.2017.2715284 (cited on page 45).

[21] L.-C. Chen et al. "DeepLab Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4 (2018), pp. 834–848. DOI: 10.1109/TPAMI.2017.2699184 (cited on page 1).

[22] Y. Chen et al. "Thoracic Low-Dose CT Image Processing using an Artifact Suppressed Large-Scale Nonlocal Means." In: *Physics in Medicine and Biology* 57.9 (Apr. 2012), pp. 2667–2688. DOI: 10.1088/0031-9155/57/9/2667 (cited on page 45).

[23] Z. Chen et al. "A limited-angle CT reconstruction method based on anisotropic TV minimization." In: *Physics in Medicine and Biology* 58.7 (Mar. 2013), pp. 2119–2141. DOI: 10.1088/0031-9155/58/7/2119 (cited on page 2).

[24] M. Defrise and R. Clack. "A Cone-Beam Reconstruction Algorithm using Shift-Variant Filtering and Cone-Beam Backprojection." In: *IEEE Transactions on Medical Imaging* 13.1 (1994), pp. 186–195. DOI: 10.1109/42.276157 (cited on page 40).

[25] J. Domke. "Generic Methods for Optimization-Based Modeling." In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by N. D. Lawrence and M. Girolami. Vol. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands: PMLR, Apr. 2012, pp. 318–326 (cited on page 51).

[26] D. L. Donoho. "Compressed Sensing." In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306. DOI: 10.1109/TIT.2006.871582 (cited on pages 42, 67).

[27] Y. Du and I. Mordatch. "Implicit Generation and Modeling with Energy Based Models." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019 (cited on pages 54, 55).

[28] Y. Du et al. "Improved Contrastive Divergence Training of Energy Based Models." In: *CoRR* abs/2012.01316 (2020) (cited on page 54).

[29] S. Duane et al. "Hybrid Monte Carlo." In: *Physics Letters B* 195.2 (1987), pp. 216–222. ISSN: 0370-2693. DOI: 10.1016/0370-2693(87)91197-X (cited on page 54).

[30] L. A. Feldkamp, L. C. Davis, and J. W. Kress. "Practical Cone-Beam Algorithm." In: *Journal of the Optical Society of America A* 1.6 (June 1984), pp. 612–619. DOI: 10.1364/JOSAA.1.000612 (cited on page 40).

[31] J. Fessler and S. Booth. "Conjugate-gradient preconditioning methods for shift-variant PET image reconstruction." In: *IEEE Transactions on Image Processing* 8.5 (1999), pp. 688–699. DOI: 10.1109/83.760336 (cited on page 33).

[32] S. Geman and D. Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6 (1984), pp. 721–741. DOI: 11.1109/TPAMI.1984.4767596 (cited on pages 53, 54).

[33] P. Getreuer. "Rudin-Osher-Fatemi Total Variation Denoising using Split Bregman." In: *Image Processing On Line* 2 (2012), pp. 74–95. DOI: 10.5201/ipol.2012.g-tvd (cited on page 2).

[34] M. U. Ghani and W. C. Karl. "CNN based Sinogram Denoising for Low-Dose CT." In: *Imaging and Applied Optics*. Optical Society of America, 2018, p. MM2D.5. DOI: 10.1364/MATH.2018.MM2D.5 (cited on page 44).

[35] I. J. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016 (cited on page 1).

[36] I. J. Goodfellow et al. "Generative Adversarial Nets." In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 2672–2680 (cited on page 54).

[37] P. Grangeat. "Mathematical Framework of Cone-Beam 3D Reconstruction via the First Derivative of the Radon Transform." In: *Mathematical Methods in Tomography*. Ed. by G. T. Herman, A. K. Louis, and F. Natterer. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991, pp. 66–97. ISBN: 978-3-540-46615-4 (cited on page 40).

[38] U. Grenander and M. I. Miller. "Representations of Knowledge in Complex Systems." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 56.4 (1994), pp. 549–603. ISSN: 00359246 (cited on page 54).

[39] S. Ha and K. Mueller. "A Look-Up Table-Based Ray Integration Framework for 2-D/3-D Forward and Back Projection in X-Ray CT." In: *IEEE Transactions on Medical Imaging* PP (Aug. 2017), pp. 1–1. DOI: 10.1109/TMI.2017.2741781 (cited on page 29).

[40] K. Hammernik et al. "A Deep Learning Architecture for Limited-Angle Computed Tomography Reconstruction." In: *Bildverarbeitung für die Medizin 2017*. Ed. by K. H. Maier-Hein geb. Fritzsche et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2017, pp. 92–97. ISBN: 978-3-662-54345-0 (cited on page 2).

[41] K. Hammernik et al. "Learning a variational network for reconstruction of accelerated MRI data." In: *Magnetic Resonance in Medicine* 79.6 (Nov. 2017), pp. 3055–3071. DOI: 10.1002/mrm.26977 (cited on page 2).

[42] W. K. Hastings. "Monte Carlo Sampling Methods using Markov Chains and Their Applications." In: *Biometrika* 57.1 (Apr. 1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97 (cited on page 54).

[43] G. E. Hinton. "Training Products of Experts by Minimizing Contrastive Divergence." In: *Neural Computation* 14.8 (Aug. 2002), pp. 1771–1800. DOI: 10.1162/089976602760128018 (cited on page 53).

[44] J. Hsieh. "Adaptive Streak Artifact Reduction in Computed Tomography Resulting from Excessive X-Ray Photon Noise." In: *Medical Physics* 25.11 (Oct. 1998), pp. 2139–2147. DOI: 10.1118/1.598410 (cited on page 44).

[45] C. Jiao et al. "Multiscale Noise Reduction on Low-Dose CT Sinogram by Stationary Wavelet Transform." In: *IEEE Nuclear Science Symposium Conference Record*. 2008, pp. 5339–5344. DOI: 10.1109/NSSMIC.2008.4774439 (cited on page 44).

[46] H. E. Johns and J. R. Cunningham. *The Physics of Radiology*. Charles C. Thomas Publisher, Limited, 2014. ISBN: 9780398090166 (cited on page 11).

[47] M. Kachelrieß, O. Watzke, and W. A. Kalender. "Generalized Multi-Dimensional Adaptive Filtering for Conventional and Spiral Single-Slice, Multi-Slice, and Cone-Beam CT." In: *Medical Physics* 28.4 (Apr. 2001), pp. 475–490. DOI: 10.1118/1.1358303 (cited on page 44).

[48] S. Kaczmarz. "Angenäherte Auflösung von Systemen linearer Gleichungen." In: *Bulletin International de l'Académie Polonaise des Sciences et des Lettres A* (1937), pp. 355–357 (cited on page 31).

[49] K. Kamnitsas et al. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation." In: *Medical Image Analysis* 36 (2017), pp. 61–78. ISSN: 1361-8415. DOI: 10.1016/j.media.2016.10.004 (cited on page 1).

[50] C. Kamphuis and F. Beekman. "Accelerated iterative transmission CT reconstruction using an ordered subsets convex algorithm." In: *IEEE Transactions on Medical Imaging* 17.6 (1998), pp. 1101–1105. DOI: 10.1109/42.746730 (cited on page 33).

[51] D. Kang et al. "Image Denoising of Low-Radiation Dose Coronary CT Angiography by an Adaptive Block-matching 3D Algorithm." In: *Medical Imaging 2013: Image Processing*. Ed. by S. Ourselin and D. R. Haynor. SPIE, Mar. 2013. DOI: 10.1117/12.2006907 (cited on page 45).

[52] D. Karimi and R. Ward. "A Denoising Algorithm for Projection Measurements in Cone-Beam Computed Tomography." In: *Computers in Biology and Medicine* 69 (2016), pp. 71–82. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2015.12.007 (cited on page 44).

[53] D. Kim and J. A. Fessler. "Accelerated ordered-subsets algorithm based on separable quadratic surrogates for regularized image reconstruction in X-ray CT." In: *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2011, pp. 1134–1137. DOI: 10.1109/ISBI.2011.5872601 (cited on page 33).

[54] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization." In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2015 (cited on page 58).

[55] O. Klein and Y. Nishina. "On the Scattering of Radiation by Free Electrons According to Dirac's New Relativistic Quantum Dynamics." In: *The Oskar Klein Memorial Lectures*. 1994, pp. 113–129. DOI: 10.1142/9789814335911_0006 (cited on page 11).

[56] E. Kobler et al. "Total Deep Variation for Linear Inverse Problems." In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2020 (cited on pages 2, 62).

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012 (cited on page 1).

[58] H. Kudo and T. Saito. "Derivation and Implementation of a Cone-Beam Reconstruction Algorithm for Nonplanar Orbits." In: *IEEE Transactions on Medical Imaging* 13.1 (1994), pp. 196–211. DOI: 10.1109/42.276158 (cited on page 40).

[59] K. Kunisch and T. Pock. "A Bilevel Optimization Approach for Parameter Learning in Variational Models." In: *SIAM Journal on Imaging Sciences* 6.2 (Jan. 2013), pp. 938–983. DOI: 10.1137/120882706 (cited on page 51).

[60] Z. Li et al. "Adaptive Nonlocal Means Filtering based on Local Noise Level for CT Denoising." In: *Medical Physics* 41.1 (Dec. 2013), p. 011908. DOI: 10.1118/1.4851635 (cited on page 45).

[61] Y. Liu and Y. Zhang. "Low-dose CT restoration via stacked sparse denoising autoencoders." In: *Neurocomputing* 284 (2018), pp. 80–89. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2018.01.015 (cited on page 45).

[62] Y. Liu et al. "Total Variation-Stokes Strategy for Sparse-View X-Ray CT Image Reconstruction." In: *IEEE Transactions on Medical Imaging* 33.3 (Mar. 2014), pp. 749–763. DOI: 10.1109/tmi.2013.2295738 (cited on pages 2, 49).

[63] Y. Liu et al. "Detecting Cancer Metastases on Gigapixel Pathology Images." In: *CoRR* abs/1703.02442 (2017) (cited on page 1).

[64] J. Ma et al. "Low-Dose Computed Tomography Image Restoration using Previous Normal-Dose Scan." In: *Medical Physics* 38.10 (Sept. 2011), pp. 5713–5731. DOI: 10.1118/1.3638125 (cited on page 45).

[65] M. Magnusson. *Linogram and Other Direct Fourier Methods for Tomographic Reconstruction*. 1993 (cited on page 25).

[66] S. Mehmood and P. Ochs. "Automatic Differentiation of Some First-Order Methods in Parametric Optimization." In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, Aug. 2020, pp. 1584–1594 (cited on page 52).

[67] T. R. Moen et al. "Low-dose CT image and projection dataset." In: *Medical Physics* 48.2 (2021), pp. 902–911. DOI: 10.1002/mp.14594 (cited on page 58).

[68] H. Morneburg and Siemens Aktiengesellschaft. *Bildgebende Systeme für die medizinische Diagnostik: Röntgendiagnostik und Angiographie, Computertomographie, Nuklearmedizin, Magnetresonanztomographie, Sonographie, integrierte Informationssysteme*. Publicis MCD Verlag, 1995. ISBN: 9783895780028 (cited on page 39).

[69] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012. ISBN: 9780262018029 (cited on page 51).

[70] S. Nah, T. H. Kim, and K. M. Lee. "Deep Multi-Scale Convolutional Neural Network for Dynamic Scene Deblurring." In: *IEEE Conference on Computer Vision and Pattern Recognition*. July 2017 (cited on page 1).

[71] National Council of Radiation Protection and Measurements. *Ionising radiation exposure of the population of the United States*. OMICS Publishing Group, Aug. 2009 (cited on pages 1, 41).

[72] E. Nijkamp et al. "Learning Non-Convergent Non-Persistent Short-Run MCMC Toward Energy-Based Model." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019 (cited on page 54).

[73] E. Nijkamp et al. "On the Anatomy of MCMC-Based Maximum Likelihood Learning of Energy-Based Models." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04 (Apr. 2020), pp. 5272–5280. DOI: 10.1609/aaai.v34i04.5973 (cited on pages 54, 55, 57, 59).

[74] S. Niu et al. "Sparse-View X-Ray CT Reconstruction via Total Generalized Variation Regularization." In: *Physics in Medicine and Biology* 59.12 (May 2014), pp. 2997–3017. DOI: 10.1088/0031-9155/59/12/2997 (cited on page 49).

[75] G. Nolet. "Seismic Wave Propagation and Seismic Tomography." In: *Seismic Tomography: With Applications in Global Seismology and Exploration Geophysics*. Ed. by G. Nolet. Dordrecht: Springer Netherlands, 1987, pp. 1–23. ISBN: 978-94-009-3899-1. DOI: 10.1007/978-94-009-3899-1_1 (cited on page 6).

[76] J. L. Prince. *Medical Imaging: Signals And Systems*. eng. 2nd ed. 2015. ISBN: 9780132145183 (cited on page 6).

[77] J. Radon. "Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten." In: *Berichte über die Verhandlungen der Königlich-Sächsischen Akademie der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse* 69 (1917), pp. 262–277 (cited on page 20).

[78] G. N. Ramachandran and A. V. Lakshminarayanan. "Three-dimensional Reconstruction from Radiographs and Electron Micrographs: Application of Convolutions instead of Fourier Transforms." In: *Proceedings of the National Academy of Sciences* 68.9 (1971), pp. 2236–2240. ISSN: 0027-8424. DOI: 10.1073/pnas.68.9.2236 (cited on page 28).

[79] S. Ramani and J. A. Fessler. "A Splitting-Based Iterative Algorithm for Accelerated Statistical X-Ray CT Reconstruction." In: *IEEE Transactions on Medical Imaging* 31.3 (2012), pp. 677–688. DOI: 10.1109/TMI.2011.2175233 (cited on page 33).

[80] S. Rit et al. "On-the-fly Motion-Compensated Cone-Beam CT Using an A-priori Model of the Respiratory Motion." In: *Medical Physics* 36.6Part1 (2009), pp. 2283–2296. DOI: 10.1118/1.3115691 (cited on page 40).

[81] G. O. Roberts and J. S. Rosenthal. "Optimal scaling of discrete approximations to Langevin diffusions." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.1 (Feb. 1998), pp. 255–268. DOI: 10.1111/1467-9868.00123 (cited on page 54).

[82] G. O. Roberts and R. L. Tweedie. "Exponential Convergence of Langevin Distributions and their Discrete Approximations." In: *Bernoulli* 2.4 (1996), pp. 341–363. DOI: bj/1178291835 (cited on pages 54, 55).

[83] P. J. Rossky, J. D. Doll, and H. L. Friedman. "Brownian dynamics as smart Monte Carlo Simulation." In: *The Journal of Chemical Physics* 69.10 (1978), pp. 4628–4633. DOI: 10.1063/1.436415 (cited on page 54).

[84] S. Roth and M. J. Black. "Fields of Experts: A Framework for Learning Image Priors." In: *IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2. 2005, 860–867 vol. 2. DOI: 10.1109/CVPR.2005.160 (cited on pages 50, 53, 62).

[85] L. I. Rudin, S. Osher, and E. Fatemi. "Nonlinear total variation based noise removal algorithms." In: *Physica D: Nonlinear Phenomena* 60.1 (1992), pp. 259–268. ISSN: 0167-2789. DOI: 10.1016/0167-2789(92)90242-F (cited on pages 1, 2).

[86] Y. Saad and H. A. van der Vorst. "Iterative solution of linear systems in the 20th century." In: *Journal of Computational and Applied Mathematics* 123.1 (2000). Numerical Analysis 2000. Vol. III: Linear Algebra, pp. 1–33. ISSN: 0377-0427. DOI: 10.1016/S0377-0427(00)00412-X (cited on page 1).

[87] K. G. G. Samuel and M. F. Tappen. "Learning Optimized MAP Estimates in Continuously-Valued MRF Models." In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 477–484. DOI: 10.1109/CVPR.2009.5206774 (cited on page 51).

[88] H. Schumacher and B. Fischer. "A New Flexible Reconstruction Framework for Motion Correction in SPECT Imaging." In: *IEEE Transactions on Nuclear Science* 54 (2007), pp. 480–485 (cited on page 40).

[89] L. A. Shepp and B. F. Logan. "The Fourier reconstruction of a head section." In: *IEEE Transactions on Nuclear Science* 21.3 (1974), pp. 21–43. DOI: 11.1109/TNS.1974.6499235 (cited on page 34).

[90] E. Y. Sidky, C.-M. Kao, and X. Pan. "Accurate Image Reconstruction from Few-Views and Limited-Angle Data in Divergent-Beam CT." In: *Journal of X-Ray Science and Technology* 14 (2006), pp. 119–139 (cited on pages 49, 67).

[91] E. Y. Sidky and X. Pan. "Image Reconstruction in Circular Cone-Beam Computed Tomography by Constrained, Total-Variation Minimization." In: *Physics in Medicine and Biology* 53.17 (Aug. 2008), pp. 4777–4807. DOI: 10.1088/0031-9155/53/17/021 (cited on pages 49, 67).

[92] J. H. Siewerdsen et al. "The Influence of Antiscatter Grids on Soft-Tissue Detectability in Cone-Beam Computed Tomography with Flat-Panel Detectors." In: *Medical Physics* 31.12 (2004), pp. 3506–3520. DOI: 10.1118/1.1819789 (cited on page 39).

[93] J. Tang, B. E. Nett, and G.-H. Chen. "Performance Comparison between Total Variation (TV)-Based Compressed Sensing and Statistical Iterative Reconstruction Algorithms." In: *Physics in Medicine and Biology* 54.19 (Sept. 2009), pp. 5781–5804. DOI: 10.1088/0031-9155/54/19/008 (cited on page 49).

[94] J.-B. Thibault et al. "A Three-Dimensional Statistical Approach to Improved Image Quality for Multislice Helical CT." In: *Medical Physics* 34.11 (Oct. 2007), pp. 4526–4544. DOI: 10.1118/1.2789499 (cited on page 44).

[95] Z. Tian et al. "Low-Dose CT Reconstruction via Edge-Preserving Total Variation Regularization." In: *Physics in Medicine and Biology* 56.18 (Aug. 2011), pp. 5949–5967. DOI: 10.1088/0031-9155/56/18/011 (cited on page 49).

[96] T. Tieleman. "Training Restricted Boltzmann Machines Using Approximations to the Likelihood Gradient." In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland: Association for Computing Machinery, 2008, pp. 1064–1071. ISBN: 9781605582054. DOI: 10.1145/1390156.1390290 (cited on page 58).

[97] A. N. Tikhonov. "On the solution of ill-posed problems and the method of regularization." In: *Doklady Akademii Nauk SSSR* 151 (1963), pp. 501–504. ISSN: 0002-3264 (cited on page 47).

[98] G. Wang, M. W. Vannier, and P.-C. Cheng. "Iterative X-ray Cone-Beam Tomography for Metal Artifact Reduction and Local Region Reconstruction." In: *Microscopy and Microanalysis* 5.1 (1999), pp. 58–65. DOI: 10.1017/S1431927699000057 (cited on page 1).

[99] A. B. Wolbarst. *Physics of Radiology*. Medical Physics Pub., 2005. ISBN: 9781930524224 (cited on page 11).

[100] J. Yang et al. "High-Order Total Variation Minimization for Interior Tomography." In: *Inverse Problems* 26.3 (Feb. 2010), p. 035013. DOI: 10.1088/0266-5611/26/3/035013 (cited on page 49).

[101] W. Yang et al. "Improving Low-Dose CT Image Using Residual Convolutional Network." In: *IEEE Access* 5 (2017), pp. 24698–24705. DOI: 10.1109/access.2017.2766438 (cited on page 45).

[102] L. Yu et al. "Radiation Dose Reduction in Computed Tomography: Techniques and Future Perspective." In: *Imaging in Medicine* 1.1 (Oct. 2009), pp. 65–84. DOI: 10.2217/iim.09.5 (cited on pages 1, 41).

[103] K. Zhang et al. "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising." In: *IEEE Transactions on Image Processing* 26.7 (2017), pp. 3142–3155 (cited on page 1).

[104] R. Zhang. "Making Convolutional Networks Shift-Invariant Again." In: *International Conference on Machine Learning*. 2019 (cited on page 57).

[105] Y. Zhang et al. "Few-View Image Reconstruction combining Total Variation and a High-Order Norm." In: *International Journal of Imaging Systems and Technology* 23.3 (Aug. 2013), pp. 249–255. DOI: 10.1002/ima.22058 (cited on pages 2, 49).

[106] T. Zhao, M. McNitt-Gray, and D. Ruan. "A convolutional neural network for ultra-low-dose CT denoising and emphysema screening." In: *Medical Physics* 46.9 (July 2019), pp. 3941–3950. DOI: 10.1002/mp.13666 (cited on page 45).

[107] B. Zhu et al. "Image Reconstruction by Domain-Transform Manifold Learning." In: *Nature* 555.7697 (Mar. 2018), pp. 487–492. DOI: 10.1038/nature25988 (cited on pages 46, 47).

[108] L. Zhu, N. R. Bennett, and R. Fahrig. "Scatter Correction Method for X-Ray CT Using Primary Modulation: Theory and Preliminary Results." In: *IEEE Transactions on Medical Imaging* 25.12 (2006), pp. 1573–1587. DOI: 10.1109/TMI.2006.884636 (cited on page 41).

[109] S. C. Zhu, Y. Wu, and D. Mumford. "Filters, Random Fields and Maximum Entropy (FRAME): Towards a Unified Theory for Texture Modeling." In: *International Journal of Computer Vision* 27.2 (1998), pp. 107–126. DOI: 10.1023/a:1007925832420 (cited on pages 50, 52).