Algorithms for Biomedical Image Reconstruction
A Brief Tutorial

Michael T. McCann   Michael Unser

École polytechnique fédérale de Lausanne
Abstract

In this tutorial, we provide a unified introduction to biomedical image reconstruction, including system modeling and reconstruction via direct, variational, and learning-based methods. Imaging is a critical tool in biological research and medicine, and most imaging systems necessarily use an image-reconstruction algorithm to create an image. While these algorithms have been developed since at least the 1960s, the field has not converged to a single, universally-accepted approach to image reconstruction. In fact, algorithms both old and new have practical applications, and new algorithms often make use of old ones in surprising ways. So, while our introduction to this topic is organized chronologically, our goal is not to provide history; rather, we aim to introduce a toolbox of ideas and algorithms that are useful for working on modern image reconstruction problems.
List of Abbreviations

**ADMM** alternating direction method of multipliers
**CCD** charge-coupled device
**CG** conjugate gradient
**CNN** convolutional neural network
**CT** computed tomography
**DCT** discrete cosine transform
**ET** electron tomography
**FBP** filtered back projection
**FFT** fast Fourier transform
**i.i.d.** independent and identically distributed
**ISTA** iterative shrinkage and thresholding
**MAP** maximum a posteriori
**MMSE** minimum mean square error
**MRI** magnetic resonance imaging
**MSE** mean squared error
**PDF** probability distribution function
**PET** positron emission tomography
**PSF** point spread function
**SIM** structured-illumination microscopy
**SNR** signal-to-noise ratio

**SPECT** single-photon emission computed tomography

**SSIM** structural similarity index

**TCIA** The Cancer Imaging Archive

**TV** total variation

**USC-SIPI** University of Southern California Signal and Image Processing Institute
Chapter 1

Introduction

Biomedical imaging is a huge and diverse field: there are a plethora of imaging devices using, e.g., light, X-rays, sound waves, magnetic fields, electrons, or protons, to measure structures ranging from nano- to macroscale. In many cases, computer software is needed to turn the signals collected by the hardware into a meaningful image. These computer algorithms are similarly diverse and numerous. For example, Google Scholar lists around 1,600 papers with the words ”MRI reconstruction” in the title.

In this tutorial, we aim to present a wide swath of biomedical image reconstruction algorithms inside a single framework. Our goal is not to review the totality of the biomedical image reconstruction literature, nor even to provide a comprehensive tutorial (though we have included references throughout that we hope will be useful pointers for a reader wishing to learn more). Instead, we have focused on creating a brief and coherent narrative that traverses some six decades of research. The result is necessarily a little stylized, and, of course, reflects the biases of the authors.

Considering many algorithms under a single, unifying framework is not only a didactic choice, it is also a practical one. While developing high-quality reconstruction algorithms requires a deep understanding of both the physics of the imaging system and the biomedical questions at hand, we have found great benefit in our own work from adopting a unified point of view. We model imaging modalities as combinations of a small set of building blocks, which allows us to see connections between modalities and quickly port ideas and computer code from one to the next. When it comes time to solve a reconstruction problem, the algorithms can treat the imaging model as a black box (possibly plus a handful of easily identifiable properties), meaning that one algorithm can work for many modalities. Such modular solutions may not be completely optimal, but they are often sufficient for a given application. When they are not, they still serve as a valuable jumping-off point.
The structure of the tutorial is as follows. We begin by introducing the concept of forward models and a set of building blocks for creating them in Section 2. In the next three sections we introduce reconstruction algorithms in a roughly chronological way, beginning with direct reconstruction and $\ell_2$ regularization (Section 3), followed by modern, sparsity-based techniques (Section 4), and ending with learning-based methods (Section 5). We compare these approaches and give our outlook on the next steps for the field in Section 6.
Chapter 2

Forward Models

The goal of any imaging system is recover an image, which is a map of a meaningful physical quantity that varies over space. In biomedical imaging, this might be the concentration of fluorophores in a cell or the X-ray attenuation coefficient of the tissues in a person’s body. Mathematically, we describe an image as a function, \( f \), that maps a point in space \( \mathbf{x} \in \mathbb{R}^d \) to a real number, \( y = f(\mathbf{x}) \in \mathbb{R} \). We leave the dimension, \( d \), of \( \mathbf{x} \) unspecified; it is usually between one and four: up to three spatial dimensions possibly plus time.

Figure 2.1: Examples of biomedical images. Top row: positron emission tomography (PET), computed radiography, X-ray computed tomography (CT) plus PET, X-ray CT, and microscopy. Bottom row: mammography, and contrast-enhanced X-ray CT, ultrasound, digital radiography, magnetic resonance imaging (MRI), and PET plus single-photon emission computed tomography (SPECT). (All images from the The Cancer Imaging Archive (TCIA) \[1\] except the microscopy \[2\] and ultrasound \[3\] images.)

In any practical setting, the image of interest is not directly observable, but must be reconstructed from its measurements. We view \( f \) through a physical
imaging system, $\mathcal{H}$, that maps $f$ to a vector of measurements,

$$g = \mathcal{H}\{f\} \in \mathbb{R}^M.$$ \hfill (2.1)

The measurement process encompasses all of the physics of the imaging system, including the sources of noise. It is nondeterministic, nonlinear, and impossible to perfectly characterize or simulate. The measurements, on the other hand, are necessarily discrete (indexed by integers) because they are stored digitally; they are also necessarily finite-length because computer hard drives have finite storage capacity. Biomedical imaging includes a broad diversity of imaging modalities, e.g., systems that rely on electromagnetic radiation (visible light, X-rays, gamma rays, etc.), electrons, magnetic fields, or sound waves; see Figure 2.1 for examples.

Most methods to recover an image from its measurements rely on modeling the imaging system, $\mathcal{H}$, with an operator $H$ (but not all, as we will discuss in Section 5). Developing an $H$ that is simultaneously mathematically convenient, computationally tractable, and physically accurate is a key challenge. It turns out that many biomedical imaging modalities can be approximated using combinations of just a few linear operators. In this section, we will introduce these operators along with examples.

### 2.1 Vector Spaces

Before we can describe the action of an operator, $H$, we need to specify its domain and range, i.e., where it maps from and to. Collectively, these are known as vector spaces, which are the central topic of linear algebra. For our purposes, it suffices to put the mathematical details aside and define a two special vector spaces that we will use throughout.

The first is $\mathbb{R}^N$, which is the space of finite-length lists of $N$ real numbers; this is usually what is meant when the term vector is used colloquially. Generally, we use elements of this space to represent lists of measurements or lists of coefficients. We note that complex-valued vectors can be included in this framework as well, because a vector of $N$ complex numbers can be represented by a vector of $2N$ real numbers. The second is the space of square-integrable functions $L_2(\mathbb{R}^d) = \{f : \int_{\mathbb{R}^d}|f(x)|^2dx < \infty\}$. This space is also called the space of finite-energy functions and is closely related to the concept of energy in physics. We use elements of this space to represent continuous objects, e.g. the physical object that we are trying to image. Other restrictions, (e.g. on the size of the non-zero support of a function or on its bandwidth) may also be important depending on the specifics of the imaging system; we will mention these as they arise.

---

1 Throughout the text, we use bold font for these finite-length vectors and matrices to emphasize that they are concrete objects that we can store and compute with.
2.2 Linear Operators

Our focus here is on operators that are linear.

**Definition 1.** An operator, $H : \mathcal{X} \rightarrow \mathcal{Y}$, is a linear operator when it satisfies

$$H\{af + bg\} = aH\{f\} + bH\{g\}, \quad (2.2)$$

for all scalars $a$ and $b \in \mathbb{R}$ and images $f$ and $g \in \mathcal{X}$.

Linear operators provide a good approximation for many real imaging systems. At the same time, they are mathematically convenient, e.g., they allow the reconstruction problem to be solved efficiently using convex optimization (Section 3.5). Another implication of selecting $H$ to be linear is that, by the Riesz representation theorem, each measurement can be written as the inner product between the image and a detector function

$$[g]_m = \langle \eta_m, f \rangle = \int_{\mathbb{R}^d} \eta_m(x)f(x)dx, \quad (2.3)$$

where the notation $[g]_m$ indicates the $m$th element of a vector $g$. The specific form of the $\eta_m$s is determined by the imaging system.

2.3 Building Blocks

The composition of linear operators is also linear, which means that it is useful to build up a toolbox of simple, well-understood operators from which to build more complicated ones. We begin with sampling.

**Definition 2.** The sampling operator, $S : L^2(\mathbb{R}^d) \rightarrow \mathbb{R}^M$, returns a vector of the values of $f$ at $M$ known locations,

$$[S\{f\}]_m = f(x_m). \quad (2.4)$$

See Figure 2.2 for an example.

Sampling is not a particularly useful imaging model on its own; it is useful because it allows us to express imaging as a sampled version of a well-understood continuous operation, such as the Fourier transform.

**Definition 3.** The Fourier transform, $\mathcal{F} : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$, expresses $f$ in terms of its frequency components,

$$\mathcal{F}\{f\}(\omega) = \hat{f}(\omega) = \int_{\mathbb{R}^d} f(x)e^{-j(\omega,x)}dx. \quad (2.5)$$

Here, we have introduced the common shorthand that $\hat{f}$ stands for the Fourier transform of $f$. See Figure 2.3 for an example.
CHAPTER 2. FORWARD MODELS

Figure 2.2: Sampling the flag of Seychelles, \( f \), at the three locations specified by \( \{x_m\} \) results in the vector of measurements \( g \).

Figure 2.3: Applying the Fourier transform to the flag of Seychelles, \( f \). Plotted here (in log-scale) is the magnitude of the result, which is a complex-valued function of the frequency vector \( \omega \). High values (lighter pixels) in the Fourier transform correspond to the directions of greatest change; this explains the white lines in \( \hat{f} \) perpendicular to the edges in the flag. (The Fourier transform has been computed analytically thanks to [4].)
2.3. BUILDING BLOCKS

Example 1 (Magnetic resonance imaging). The Fourier transform and sampling operators are sufficient to model MRI; this is known in the field as the $k$-space formalism \cite{5}. In MRI, the physical quantity of interest is the positron density and the measured signal is given by

$$g = (S \circ \mathcal{F})\{f\}, \quad (2.6)$$

where $\circ$ indicates operator composition, $(A \circ B)\{f\} = A\{B\{f\}\}$. The locations of the Fourier samples depend on the hardware setup. Typical choices include Cartesian sampling, random sampling, and helical sampling.

Definition 4. The inverse Fourier transform, $\mathcal{F}^{-1} : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$, recovers $f$ from its Fourier transform,

$$f(x) = \mathcal{F}^{-1}\{\hat{f}\}(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(\omega) e^{j\langle\omega, x\rangle} d\omega. \quad (2.7)$$

The fact that the Fourier transform is invertible suggests a method for MRI reconstruction: approximate the integral (2.7) with a sum over the known $[g]_m = \hat{f}(\omega_m)$. As the measurement density increases, this method becomes more and more accurate. We will explore this approach in more detail in Section 3.1.

Definition 5. The multiplication operator, $M : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$, multiplies $f$ pointwise by another image,

$$M_h\{f\}(x) = f(x)h(x), \quad (2.8)$$

where $h$ must be bounded: for all $x$, $|h(x)| < \infty$.

One example of multiplication is windowing (Figure 2.4), where $h$ is a function taking values between zero and one, such as a rect

$$\text{rect}(x) = \begin{cases} 1 & x \in [-\frac{1}{2}, \frac{1}{2}]^D \\ 0 & \text{otherwise.} \end{cases} \quad (2.9)$$

While it is mathematically convenient to work with infinite signals (e.g., note the integral over all space in the definition of the inverse Fourier transform), real imaging systems have a limited spatial extent; windowing can model this limit.

Another example of multiplication is modulation, where $h$ is a complex exponential,

$$h(x) = e^{j(\omega_0, x)}. \quad (2.10)$$

The effect of modulation is easier to see in the frequency domain,

$$(\mathcal{F} \circ M_{\omega_0}\{f\})(\omega) = \mathcal{F}\{f\}(\omega - \omega_0). \quad (2.11)$$

Thus, multiplication by a complex exponential results in a shift in the frequency content of the image.
Figure 2.4: Windowing the flag of Seychelles, $f$, by multiplication with a shifted $\text{rect}$. We can imagine this operation as modeling a limited field of view, like looking at the flag through a microscope.

**Example 2** (MRI with coil sensitivity). A more accurate model of MRI accounts for spatial variations in the coil sensitivity,

$$g = (S \circ \mathcal{F} \circ M_h)\{f\},$$

where $h$ is measured as part of calibration. For a detailed treatment of the MRI forward model, see [6].

**Definition 6.** The convolution operator, $C_h : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$, implements the linear, shift-invariant system with impulse response $h$,

$$C_h\{f\}(x) = \int_{\mathbb{R}^d} h(x - x')f(x')dx',$$

where $h$ must be absolutely integrable ($\int_{\mathbb{R}^d} |h(x)|dx < \infty$). Convolution can be equivalently defined in the frequency domain,

$$C_h\{f\}(x) = (\mathcal{F}^{-1} \circ M_{\hat{h}} \circ \mathcal{F})\{f\}(x),$$

where $\hat{h}$ is the frequency response of the system, $\hat{h} = \mathcal{F}\{h\}$.

Although the convolution is a composition of Fourier transforms and a multiplication, it deserves its own symbol because of its ubiquitous role in imaging. What makes it so useful is that many measurement systems are shift-invariant, because the underlying physics work the same everywhere and
the sensors are placed on a regular grid, e.g., pixels on a charge-coupled device (CCD) camera. Convolution is also important because discrete convolutions can be implemented using the fast Fourier transform (FFT), which can result in fast reconstruction algorithms.

**Example 3** (Brightfield microscopy). In brightfield microscopy, what appears at the eyepiece is a blurred version of what is on the specimen plane,

$$g = C_h \{f\}. \quad (2.15)$$

The blurring function is called the Airy disk,

$$h(x) = \left( \frac{2J_1(r)}{r} \right)^2, \quad r = \frac{2\pi a \|x\|}{\lambda R}, \quad (2.16)$$

where $J$ is the Bessel function of the first kind, $a$ is the radius of the aperture, $\lambda$ is the wavelength of the illumination, and $R$ is the focal length. The width of this function plays an important role in the achievable resolution—the minimum distance that two point sources must be separated in order for them to be distinguished by eye. For visible light, this limit is around 250 nm.

For another perspective of the resolution limit, we note that the Fourier transform of the Airy disk is the indicator function of a circle, i.e., for some $w_{\text{max}}$ it is one for $\|\omega\| \leq w_{\text{max}}$ and zero for $\|\omega\| > w_{\text{max}}$. Functions of this form are called low-pass functions. Writing the measurement process in the Fourier domain we have

$$F_g = (M_h \circ F) \{f\}, \quad (2.17)$$

and therefore the high-frequency content of $f$ has no effect on $g$ because everything outside of $w_{\text{max}}$ is multiplied by zero during the measurement. For more information on microscopy, see [7].

**Example 4** (Structured-illumination microscopy). The idea behind structured-illumination microscopy (SIM) is to use a special illumination pattern to modulate the image $f$, rearranging its frequency content so that its high frequencies are no longer destroyed by the measurement process. The continuous forward model is

$$g = (C_h \circ M_p) \{f\}, \quad (2.18)$$

where $p$ is a complex exponential, $p(x) = e^{j(\omega_0 \cdot x)}$. Viewed in the Fourier domain,

$$F\{g\}(\omega) = (M_h \circ F \circ M_p) \{f\}(\omega) = \hat{h}(\omega)F\{f\}(\omega - \omega_0). \quad (2.19)$$

In order to recover both the high-frequency and low-frequency content of $f$, the imaging process can be repeated with different illumination patterns, resulting in different values for the frequency shift, $\omega_0$. For more information on SIM, see [8].
Definition 7. A change of variables operator, $\Phi_\varphi : L^2(X) \to L^2(Y)$, is a linear operator such that $\Phi_\varphi f = f(\varphi(\cdot))$, where $\varphi : Y \to X$ is a diffeomorphism (a smooth function with a smooth inverse).

We have already seen one change of variables: the frequency shift caused by multiplication by a complex exponential. Using a change of variables operator, we can express a shift as

$$f(x - x_0) = \Phi_\varphi \{f\}(x), \quad \text{with} \quad \varphi(x) = x - x_0. \tag{2.20}$$

See Figure 2.5 for another example.

Figure 2.5: A change of variables applied to the flag of Seychelles, $f$. Each point is expressed in terms of its polar coordinates: the angle it makes with the $x_1$ axis and its distance from the origin.

Example 5 (X-ray computed tomography). The forward model for a variety of CT imagining modalities, including X-ray CT, PET, SPECT, and electron tomography (ET) is called the X-ray transform. It measures the line integrals of an image and is usually given in 2D by

$$X_\theta f(y) = \int_{\mathbb{R}} f(y\theta + t\theta^\perp)dt, \tag{2.21}$$

where $\theta$ is an angle, $\theta = (\cos \theta, \sin \theta)$ is a unit vector, and $\theta^\perp = (-\sin \theta, \cos \theta)$ is its perpendicular. It turns out that the X-ray transform can also be written in terms of the building blocks we have already defined as

$$(\mathcal{F}_{1D} \circ X_\theta)\{f\} = (\Phi_y \circ \mathcal{F}_{2D})\{f\}, \tag{2.22}$$

where the subscripts on the Fourier transforms remind us that that the image is two dimensional while its X-ray transform is one dimensional. We verify
2.4. DISCRETIZATION

this by writing

\( (\Phi_\theta \circ \mathcal{F}_{2D}) \{ f \} = \int_{\mathbb{R}^2} f(x) e^{-j(y \theta, x)} \, dx \) (2.23)

\[ = \int_{\mathbb{R}^2} f(Aa) e^{-j(y, Aa)} \, da \] (2.24)

\[ = \int_{\mathbb{R}^2} f(Aa) e^{-jy[a]_1} \, da \] (2.25)

\[ = \int_{\mathbb{R}} \int_{\mathbb{R}} f(Aa) d[a]_2 e^{-jy[a]_1} \, d[a]_1 \] (2.26)

\[ = \mathcal{F}_{1D} \circ X_\theta f(y), \] (2.27)

where trick is that we can always use the orthogonal matrix \( A = [\theta \quad \theta^\perp] \) to express \( x \) as \( Aa \) for some \( a \). The result is called the Fourier slice theorem or central slice theorem, because it relates the 1D Fourier transform of the X-ray transform of an image to a slice through its 2D Fourier transform. For more information see [9] for X-ray CT, with additional details on PET in [10] and ET in [11].

2.4 Discretization

We have, so far, presented a toolbox of operators that can be combined to model many biomedical imaging modalities. Generally speaking, we will need to implement these operators in software in order to solve image reconstruction problems. But, how can we do this when the definitions are in the continuous domain? That is, how can we represent \( f \in L^2(\mathbb{R}^n) \) in a computer? This is the problem of discretization.

One approach to the problem is to discretize at the end: formulate the reconstruction algorithm in the continuous domain and develop discrete versions of each of the necessary operations. Anywhere a function is needed, it is represented by a finite vector of its samples; the evaluation of the function at other points is achieved via interpolation, and integrals are replaced with sums by the trapezoidal rule, etc. We will discuss this approach further in Section 3.1.

We focus here on discretizing at the beginning via the synthesis operator.

Definition 8. The synthesis operator \( T_{\{ \beta_n \}_{n=1}^N} : \mathbb{R}^N \to L^2(\mathbb{R}^d) \), expresses a function as a sum of a finite number of basis functions,

\[ f(x) = T_{\{ \beta_n \}_{n=1}^N} f(x) = \sum_{n=1}^N [f]_n \beta_n(x), \] (2.28)
where $\beta_n \in L_2(\mathbb{R}^d)$, $f$ denotes the original function and $\mathbf{f}$ is the vector of its coefficients.

The synthesis operator allows for a simple discretization of any continuous forward model, $H : L_2(\mathbb{R}^d) \rightarrow \mathbb{R}^M$, according to $\mathbf{H} = H \circ \mathbf{T}$, where $\mathbf{H}$ is an $M \times N$ matrix, called the system matrix. The discretized imaging model is then $\mathbf{g} = \mathbf{Hf}$.

Recalling that we can view the linear measurement process as taking inner products with detector functions (2.3), we can express the elements of the system matrix as inner products between the basis functions and these detector functions,

$$\left[\mathbf{A}\right]_{m,n} = \langle \eta_m, \beta_n \rangle = \int_{\mathbb{R}^d} \eta_m(x) \beta_n(x) dx.$$ 

Practically, the basis functions are chosen to be shift-invariant, $\beta_n(x) = \beta(x - x_n)$, where the $x_n$ form a regular grid, e.g., $x_n \in \{x = Tk + x_0 : k \in [0, 1, \ldots, N_1 - 1] \times [0, 1, \ldots, N_2 - 1]\}$. They may also be multiplicatively separable,

$$\beta(x) = \beta(x_1) \beta(x_2) \cdots \beta(x_d),$$

where (in a slight abuse of notation) $\beta(x)$ is the $d$-dimensional separable basis function and $\beta(x)$ is the 1D function from which it is built.

**Example 6 (Pixel model).** Because the ultimate result of image reconstruction is a digital image, it is natural to use the square pixel as a basis function, i.e. $\beta(x) = \text{rect}(x)$, see Figure 2.6 for an example. The downside of the pixel model is, generally, its lack of smoothness: It is not differentiable, so it is not appropriate for continuous forward models involving differentiation. It is also not bandlimited: its Fourier transform is

$$\text{sinc}(x) = \frac{\sin \pi x}{\pi x}$$

and therefore may not interact well with forward models that amplify high frequencies. Despite this, the pixel model is ubiquitous in practice.

## 2.5 Summary

We have presented a set of building blocks—linear operators that can be composed to model a variety of physical measurement processes. In the following, we give a nonexhaustive list of these imaging modalities:
2.5. SUMMARY

2D or 3D tomography uses X-rays. The measurements are \{(S \circ X_{\theta_i})\{f\}\}_i, which is a set of sampled X-ray projections taken from different directions, where \(i\) indexes the projection direction. Variations include parallel, cone beam, and spiral sampling patterns.

3D microscopy uses fluorescence. The measurements are \((S \circ C_h)\{f\}\), which is a blurred, sampled version of \(f\). Variations include brightfield, confocal, and light sheet microscopy.

SIM uses fluorescence. The measurements are \{((S \circ C_h \circ M_{w_i})\{f\}\}_i, which is a set of modulated, blurred, and sampled versions of \(f\), where \(i\) indexes the modulation pattern. Variations include performing 3D reconstructions or using non-sinusoidal patterns.

PET uses gamma rays. The measurements are \{((S_{y_i} \circ X_{\theta_i})\{f\}\}_i, which, like in tomography, are sampled X-ray projections. We use the subscript \(y_i\) to indicate that the sampling positions are dependent on the projection direction, and note that there may as few as one sample per direction. Variations include time-of-flight PET.

MRI uses magnetic fields and measures radio waves. The measurements are \{((S \circ F \circ M_{w_i})\{f\}\}_i, which are a set of samples of the Fourier transform of \(f\) after weighting by several different coil sensitivities, indexed by \(i\).
are several variations in the sampling pattern, including Cartesian and radial sampling, and the method can also be used for dynamic processes, e.g., cardiac MRI.

**Optical diffraction tomography** uses coherent light. Although the problem is, in general, nonlinear, using the first Born approximation the measurements are \((S \circ F \circ M_h)(f)\), where \(h\) is related to the incident wave. Variations include holography and grating-based interferometry.

In each case above, we have stated a measurement model that goes from a continuous \(f\) to a discrete \(g\). By additionally making use of a synthesis operator, each of these can be converted to a system matrix that relates the discrete image we want to recover to the discrete measurements we have access to. This matrix plays a key role in reversing the measurement process to arrive at an estimate of the image of interest, which is the topic of the following sections.

### 2.5.1 Further Reading

Many of these building blocks are fundamental concepts in signal processing; for a textbook treatment, see the classic [13] or [14]. For more information on the modalities we have discussed, see the references listed at the end of each example. For a MATLAB software library based on this methodology, see [15].
Chapter 3

Classical Image Reconstruction

Assuming that we have an accurate, implementable forward model, $H$, and measurements $g$, our task now is to reconstruct an image. Our focus is on discretizing immediately, i.e., using a synthesis operator to create a system matrix, $H$, because doing so gives access to simple, implementable reconstruction algorithms. But, before dispensing with the operator $H$ in favor of the matrix $H$, we need to discuss direct inversion methods, which are derived in the continuum.

3.1 Direct Inversion

Putting aside sampling for the moment, we can think of most imaging modalities as invertible. Abstractly, physics runs the same forwards as backwards, meaning that all the information about the image exists in the measurements. As a concrete example, note the Fourier transform is invertible. If MRI is based on the Fourier transform, it seems that the inverse Fourier transform should provide an MRI reconstruction technique. To develop a direct inversion algorithm, we formulate a forward model in the continuum (i.e., ignoring sampling), $H : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$ and derive (analytically) its inverse, $H^{-1}$. Such an inverse will usually involve integrals of a continuous version of the measurements. To actually implement $H^{-1}$, we replace these integrals with sums over the known measurements and use interpolation whenever we need to evaluate a quantity between its samples.

Example 7 (Filtered back projection). The canonical example of direct inversion is the filtered back projection (FBP) algorithm for X-ray CT, which
CHAPTER 3. CLASSICAL IMAGE RECONSTRUCTION

dates back to 1971 [16]. Starting with the X-ray transform as a forward model, we use the Fourier slice theorem to derive the following inverse

\[
H^{-1}\{g\}(x) = \int_0^\pi F_{1D}^{-1} M|\omega| F_{1D} X_\theta f (\theta, x) \, d\theta.
\]

(3.1)

This equation says to take each projection (from 0 to \(\pi\)), convolve by a filter with impulse response \(|\omega|\), and back project it (smear it across the image). The name filtered back projection is therefore a bit misleading: the filtering occurs before the back projection.

The formula (3.1) looks simple, but several approximations are needed to implement it from discrete measurements. First, the integral over \(\theta\) must be replaced with a sum. Second, the filter must be implemented digitally, with careful handling of the boundary, high frequencies (because they are amplified), and zero frequency (because it is removed). Finally, the filtered projections must be interpolated in order to back project them. Despite these approximations, the FBP is a robust and high-quality method, especially when the number of projections is large.

3.2 Variational Methods

In order to avoid the challenges associated with discretizing a direct inversion method, we prefer to work with a linear forward model that takes the form of a matrix, \(H\). In this setting, the simplest approach to image reconstruction would be to solve

\[
Hf = g
\]

(3.2)

for \(f\). The first problem with this formulation is that \(H\) is rarely invertible—in most scenarios, we want to get as many image pixels, \(N\), out of as few measurements, \(M\), as possible, leading to a short, fat forward model with no left inverse. Practically, this means that there will be many \(f\)s that satisfy (3.2). If any of the measurement vectors are linearly dependent (which can happen in any setting, but is guaranteed when \(M > N\)), any amount of measurement noise will be enough to move \(g\) out of the span of \(H\), leading to (3.2) having no solution. Finally, even when a unique solution does exist \((M = N)\), the solution is usually very sensitive to noise, to the point were direct inversion of \(H\) is useless (see Figure 3.1 for an example). Thus, (3.2) is an example of an ill-posed problem.

Definition 9. A well-posed problem satisfies:

1. A solution exists.

2. The solution is unique.
3. The solution is a continuous function of the measurements.

A problem that is not well-posed is called ill-posed.

![Image showing the effect of measurement noise on direct inversion of discrete convolution.](image)

Figure 3.1: Effect of measurement noise on direct inversion of discrete convolution. Each pixel of the measurement (b) is a local average of pixels in the image (a). This operation is invertible (c), but even a tiny amount of noise (d) is amplified by the inversion, leading to heavy artifacts in the reconstruction (e). (Image from the University of Southern California Signal and Image Processing Institute (USC-SIPI) Image Database.)

A more promising approach, which is the focus of this section and the next, is to formulate reconstruction as a minimization problem. These approaches are sometimes called variational methods; while this term has precise meanings in fields such as differential equations, quantum mechanics, and Bayesian inference, in our context it only means we aim to minimize something. As a first example, we can seek a reconstruction that minimizes the disagreement with the measurements according to

\[
\tilde{f} = \arg \min_{f} \|g - Hf\|^2.
\]  

(3.3)
From a geometric perspective, we see that this problem must have at least one solution: we can project \( g \) onto the span of \( H \), resulting in a denoised measurement, \( g' \) and solve \( Hf = g' \). However, the problem may still be ill-posed. Equating the derivative of the objective function in (3.3) with zero results in the so-called normal equation,

\[
H^T g = H^T H f,
\]

which \( f \) must satisfy to be a solution. When the number of linearly-independent measurements (rank of \( H \)) is at least the number of unknowns (length of \( f \)), then there is a unique solution \( (H^T H)^{-1} H^T g = f \). Otherwise, there are infinitely-many solutions. Specifically, the solutions form an affine subspace of \( \mathbb{R}^N \), \( N(H) + \tilde{f} \), where \( N(H) = \{ f : Hf = 0 \} \) is called the null space of \( H \) and \( \tilde{f} \) is one solution of (3.3).

**Remark 1** (Does the square matter?). One challenge in understanding the biomedical reconstruction literature is the plethora of similar-looking problem statements that may or may not be equivalent. The first such example is

\[
\arg \min_f \| g - Hf \| \quad \text{versus} \quad \arg \min_f \| g - Hf \|^2.
\]

(3.5)

These expressions are equivalent because the function \( \cdot^2 \) is monotonically increasing over the nonnegative real numbers. We will see a less-trivial generalization of this idea in Remark 2.

The version without the square is more natural: it gives a reconstruction that agrees with the measurements in the sense of Euclidean distance. The benefit of the square is that it makes the expression linear (because it cancels the square root in the definition of the norm) and leads to a simpler expression for the gradient.

The infinitely-many solutions case is the norm. This is, again, because we want the highest possible resolution, \( N \), from a fixed number of measurements, \( M \), determined by the hardware. As a result, the number of unknowns is larger than the number of measurements. Thus, (3.3) is also an ill-posed problem.

**Example 8** (Null space). Consider a toy problem where

\[
H = \begin{bmatrix}
1 & 0 & 1 \\
0 & 1 & -1 \\
1 & 1 & 0
\end{bmatrix}, \quad \text{and} \quad f^* = \begin{bmatrix}
1 \\
1 \\
2
\end{bmatrix}.
\]

(3.6)

The true measurement process is \( g = Hf^* + n = [3 \quad -1 \quad 2.1]^T \) and we want to reconstruct by solving (3.3). If we plug the problem into MATLAB, the result is \( f \approx [1.70 \quad 0.37 \quad 1.33]^T \), giving a sum of squared errors of 0.0033.
3.3. TIKHONOV REGULARIZATION

But, by changing the initialization of the solver, infinitely many equally good solutions can be obtained:

| $f_0$ | $\hat{f}$ | $\|g - Hf\|^2$ |
|-------|-----------|----------------|
| $[0 \ 0 \ 0]^T$ | $[1.70 \ 0.37 \ 1.33]^T$ | .0033 |
| $[0 \ 0 \ 1]^T$ | $[1.37 \ 0.70 \ 1.67]^T$ | .0033 |
| $[13 \ 8 \ 18]^T$ | $[-2.30 \ 4.37 \ 5.33]^T$ | .0033 |

This effect occurs because $H$ has a non-zero null space, specifically $H \begin{bmatrix} 1 & -1 & -1 \end{bmatrix}^T = 0$; note that each solution can be obtained from another by adding a scaled version of this null space vector.

3.3 Tikhonov regularization

When a formulation is ill-posed in the sense of having no solution, it is a dead end. But, when a formulation merely has infinitely-many solutions, all we need is a mechanism to pick one of them to obtain a well-posed problem; Tikhonov regularization provides a mechanism to do this.

The idea is to introduce a regularization functional, $\|Lf\|^2$, where $L$ is a linear operator, that measures the roughness of $f$ (or some other undesirable property). The reconstruction can then be formulated as

$$\arg\min_f \|Lf\|^2 \text{ subject to } \|g - Hf\|^2 \leq \sigma^2,$$

where $\sigma$ is a scalar constant selected based on the expected level of noise. An equivalent formulation is

$$\arg\min_f \|g - Hf\|^2 + \lambda \|Lf\|^2$$

where $\lambda$ is a scalar constant controlling the strength of the regularization (see Figure 3.2).

**Remark 2** (Does the square matter now?). We saw in Remark 1 that the square in the least squares formulation is only a mathematical convenience. Now, with the addition of the regularization term, we can compare a more general case,

$$\arg\min_f J_1(f) + \lambda_L J_2(f) \text{ versus } \arg\min_f \Psi_1(J_1(f)) + \lambda_R \Psi_2(J_2(f)),$$

(3.9)
Figure 3.2: Illustration of the effect of changing the Tikhonov regularization parameter, $\lambda$, when the regularization operator is a high-pass filter. Stronger regularization reduces noise, but also reduces sharpness; this can be thought of as a kind of bias–variance tradeoff.

where $J_1$ and $J_2$ are positive, differentiable functionals, $\lambda_L$ and $\lambda_R$ are positive numbers, and $\Psi_1$ and $\Psi_2$ are monotonically increasing, differentiable functions. Intuitively, these are different problems when $\lambda_L = \lambda_R$ because the $\Psi$s will change the relative importance of $J_1$ and $J_2$ (which we can think of as representing the data and regularization terms, respectively). But, if we allow a different value of $\lambda$ in each problem, they are equivalent. To see this, take a solution of the left problem, $f^*$. At this point, the gradients of $J_1$ and $\lambda_L J_2$ are equal and opposite. In the right problem, at $f^*$, the chain rule tells us that the gradient of the $J_1$ term has been multiplied by $\Psi_1'(J_1(f^*))$ and the gradient of the $J_2$ term has been multiplied by $\Psi_2'(J_2(f^*))$. Because both of these multipliers are positive, we can set

$$
\lambda_R = \frac{\Psi_1'(J_1(f^*)) \lambda_L}{\Psi_2'(J_2(f^*))}, \quad (3.10)
$$

which makes $f^*$ a critical point for the right problem. We can make the same argument in the reverse direction, and a similar argument can be made with nondifferentiable functions by using subgradients.

In practice, we set $\lambda$ by hand to give the best-looking reconstruction; this argument tells us that if we can find a good $\lambda$ for the left formulation, we will
be able to find a good \( \lambda \) in the right formulation. As a result, we have a lot of flexibility to add or remove, e.g., squares, from data fidelity and regularization terms without affecting the solutions we get.

In the Tikhonov formulation, the normal equation becomes

\[
H^T g = (H^T H + \lambda L^T L) f. \tag{3.11}
\]

This expression reveals the main advantage of the formulation: the problem will have a unique solution when the intersection of the null spaces of \( H \) and \( L \) is zero. One easy way to achieve this is to let \( L \) be full rank, e.g., when \( L \) is the identity matrix, the solution is unique and the regularization penalizes large values in \( f \).

**Remark 3** (Constraints versus regularization). We have now seen (and will continue to see) data and regularization terms on \( f \) expressed as both functionals (also called penalties), e.g., \( \| L f \| \), and as constraints, e.g., \( \| g - H f \| \leq K \). The natural question is to compare

\[
\arg \min_f J_1(f) + \lambda J_2(f) \quad \text{versus} \quad \arg \min_f J_1(f) \quad \text{s.t.} \quad J_2(f) \leq \sigma, \tag{3.12}
\]

where \( J_1 \) and \( J_2 \) are positive functionals and \( \lambda \) and \( \sigma \) are positive numbers. We will call these the penalized and constrained formulations; the first is also called Tikhonov regularization and the second Ivanov or Morozov regularization, depending on whether \( J_1 \) is the data term (Ivanov) or the regularization term (Morozov). We can afford to be loose in interchanging these forms because, roughly, solutions of one are also solutions of the other (when the \( \lambda \) and \( \sigma \) are chosen correctly). We make this notion precise in the next two theorems; for a discussion of the engineering implications, skip ahead.

**Theorem 1.** For every choice of \( \lambda \), for each solution of the penalized problem, there is a choice of \( \sigma \) that makes it a solution of the constrained problem.

**Proof.** Fix \( \lambda \) and assume \( f^P \) is a solution to the penalized problem. Set \( \sigma = J_2(f^P) \). Now for sake of contradiction, assume there exists \( f^C \), that is a better solution to the constrained problem than \( f^P \), i.e., \( J_1(f^C) < J_1(f^P) \) and \( J_2(f^C) \leq \sigma \). Then we have \( J_1(f^R) + \lambda J_2(f^R) < J_1(f^L) + \lambda J_2(f^L) \), which contradicts \( f^P \) being a solution to the penalized problem. Thus, there is no such \( f^C \), which means \( f^P \) is a solution to the constrained problem. \( \square \)

**Theorem 2.** Assume that penalized problem has a solution for every \( \lambda \geq 0 \). For every choice of \( \sigma \), for each solution of the constrained problem, there is a choice of \( \lambda \) that makes it a solution of the penalized problem.
CHAPTER 3. CLASSICAL IMAGE RECONSTRUCTION

Proof. Fix $\sigma$ and assume $f^C$ is a solution to the constrained problem. Consider the function $F(\lambda) = \min_f J_1(f) + \lambda J_2(f)$. We are searching for a value, $\lambda^*$, such that $F(\lambda^*) = J_1(f^C) + \lambda^* J_2(f^C)$, because this will make $f^C$ a solution to the penalized problem. We know that, for any $\lambda^*$, $F(0) = \min_f J_1(f) \leq J_1(f^C) + \lambda^* J_2(f^C)$ and that we can pick $\lambda^\wedge$ large enough that $F(\lambda^\wedge) \geq J_1(f^C) + \lambda^* J_2(f^C)$. We can show that $F$ is continuous, thus the intermediate value theorem gives us that there exists $\lambda^*$ such that $F(\lambda^*) = J_1(f^C) + \lambda^* J_2(f^C)$, which makes $f^C$ a solution of the penalized problem.

 Practically, these problems are equivalent; theoretically, the relationship is more subtle (e.g., we have not shown that the solution sets can be made the same, only that the problems can be made to share a solution). When more is assumed about the functionals, we can say more. Generally, which form is chosen is a matter of personal preference. For a much deeper look at this issue, see \[17\].

3.4 Statistical Formulation

We now consider a statistical measurement model

$$g = Hf + n,$$ \hspace{1cm} (3.13)

where $H$ remains a deterministic measurement matrix, but where $f$ and $n$ (and therefore $g$) are random variables. If we know (or can model or estimate) the distribution of $n$, then the model (3.13) lets us determine the probability distribution function (PDF) of the measurements given a fixed image, $p(g | f)$.

To do reconstruction, we are more interested in the distribution of $f$ conditional on the measurements, which we can obtain using Bayes’ rule as

$$p(f | g) = \frac{p(g | f)p(f)}{p(g)}.$$ \hspace{1cm} (3.14)

So, for a fixed set of measurements $g$, (3.14) is a function that states how likely a given $f$ is to have given rise to $g$. It requires a model for $p(f)$, which is analogous to the choice of a regularizer in the variational formulations we already presented. Unfortunately, it also requires computation of $p(g)$, which is typically intractable (because it requires integrating over all $f$s that can create $g$). So, while (3.14) in principle gives all the information we might want in a reconstruction, we usually do not use it directly.

Using the same statistical model, we can build algorithms that seek a single image satisfying some measure of optimality. In one such approach, we
seek the minimum mean square error (MMSE) solution, which is given by

$$\text{arg min} \ E(\|\tilde{f} - f\|^2 \mid g),$$  \hspace{1cm} (3.15)$$

where here $\tilde{f}$ is the reconstruction (which depends on the measurements, $g$) and $f$ is the ground truth. One way to think of the MMSE solution is that if it is used to perform many reconstructions it will have the lowest average error among all algorithms (assuming that all the model assumptions are correct). By differentiating (3.15) and equating it with zero, we find that the MMSE is given by $E(f \mid g)$, which is just the expectation associated with (3.14). Unfortunately, it is usually difficult to build algorithms to find the MMSE solution in practice.

A different approach is to find the maximum a posteriori (MAP) solution, which is the mode of (3.14),

$$\text{arg max} \ f \ p(f \mid g) = \text{arg max} \ f \ p(g \mid f)p(f).$$ \hspace{1cm} (3.16)$$

Put another way, this is the $f$ that is most likely to have generated the measured $g$. While the MAP solution is not always optimal (see [18] for more discussion) it has the advantage of providing a statistically coherent way to account for noise and priors on the image, while also leading to optimization problems that can be efficiently solved (when $p(f)$ is chosen carefully).

As a specific example of MMSE and MAP solutions, we return to the measurement model (3.13) and consider the case where the elements of the noise $n$ are independent and identically distributed (i.i.d.) Gaussian with zero mean and variance $\sigma^2$, and $f$ is a zero-mean Gaussian process with invertible covariance matrix $E(ff^T) = C$. In this case, we can transform the MAP objective (3.16) using the negative log and arrive at

$$\text{arg min} \ f \ \frac{1}{\sigma^2}\|g - Hf\|^2 + \|C^{-\frac{1}{2}}f\|^2,$$ \hspace{1cm} (3.17)$$

which has a closed-form solution

$$(H^T H + \sigma^2 C^{-1})^{-1} H^T g.$$ \hspace{1cm} (3.18)$$

We can derive the MMSE solution from by finding the expectation of (3.14), which essentially requires some lengthy algebra involving the underlying multivariate Gaussian distributions. The result is

$$CH^T \left(HCH^T + \sigma^2 I\right)^{-1} g,$$ \hspace{1cm} (3.19)$$

which is also called the Wiener filter. It turns out that (3.18) and (3.19) are equal. (One way to verify this is to check that both satisfy the normal equation of (3.17), $H^T g = (H^T H + \sigma^2 C^{-1})f$, which is meaningful because the
solution to (3.17) is unique.) And, further, note that (3.17) is equivalent to
the Tikhonov formulation (3.8) with $L = C^{-\frac{1}{2}}$ and $\lambda = \sigma^2$; thus, the optimal
choice of $L$ is the whitening operator matched with $f$. So, for the special case
of Gaussian denoising of a Gaussian process, the MAP and MMSE solutions
are the same, and they can also be obtained from a specific variational formulation. But, for many other image reconstruction problems, this equivalence
does not hold.

### 3.5 Iterative Reconstruction

We have now seen several reconstruction problems that have a closed-form
solution. Unfortunately, these usually involve inverting the normal matrix,
$H^T H$, which is huge (its size is number of image pixels squared). So, for practical
implementations, we turn to iterative algorithms. Iterative algorithms
for solving convex problems of this form are, by now, a textbook topic [19];
we present one such algorithm here to give a sense of their form.

Given a Tikhonov reconstruction problem,

$$
\arg \min_f J(f) = \arg \min_f \|g - Hf\|^2 + \lambda \|L f\|^2,
$$

(3.20)

The gradient descent algorithm involves iterates of the form

$$
f^{(k+1)} = f^{(k)} - \gamma^{(k)} \nabla J(f^{(k)}),
$$

(3.21)

where the gradient, $\nabla J(f)$, is given by

$$
\nabla J(f) = -H^T g + (H^T H + \lambda L^T L)f
$$

(3.22)

and setting $\gamma$ to a small constant gives a workable algorithm. One advantage
of gradient descent is that convex constraints, such as the positivity of $f$,

can be enforced by projecting onto the feasible region after each iteration,
an algorithm known as projected gradient descent. Gradient descent is good
starting point for building iterative algorithms, but remember that there are
many alternatives and the best algorithm will depend on the problem at hand.

For example, the conjugate gradient (CG) algorithm solves the same problem
more efficiently and accurately, at the cost of increased memory usage [20].

**Remark 4** (Convolutional normal matrix). It turns out that the normal ma-
trix $H^T H$ appears in many iterative reconstruction schemes and is often the
computational bottleneck. The case of gradient descent (3.22) is typical: re-
construction requires a single back projection, $H^T g$, followed by one normal

operation \((H^T H + \lambda L^T L)f\) per iteration. Often, we select \(L\) to be a convolution (e.g., a high-pass filter), which means that \(L^T L\) is a convolution as well. If \(H^T H\) is a convolution, then the entire operation is a convolution. This is important because a discrete convolution can be computed using the FFT, making it much faster than usual matrix multiplication. Luckily, \(H^T H\) turns out to be a convolution for many modalities. One surprising example of this is parallel-ray X-ray CT, where \(H^T H\) is a convolution as a consequence of the Fourier slice theorem. This result has been used in cryo-EM \([21]\) and synchrotron microtomography \([22]\).

3.6 Summary

Classical image reconstruction methods fall into either direct methods or variational (regularized, iterative) methods. Direct methods are derived in the continuous domain; they are fast to apply and give good results when the number of measurements is high. Variational methods involve minimizing an objective function that usually consists of a data term and one or more regularization terms. These methods tend to be more robust to noise than direct methods, at the cost of increased computation.
Chapter 4

Sparsity-Based Image Reconstruction

Over the past decades, numerous new reconstruction methods have emerged, many of which continue to make use of the variational framework, i.e., reconstruction is performed by minimizing a cost functional. Of these, one dominant paradigm has been that of sparsity—the idea that a high dimensional image, $f$, can be represented by only a small number of nonzero coefficients. The concept of sparsity is included in reconstruction problems by changing from the quadratic regularization functional, $\|Lf\|^2$, to one that promotes sparsity. Especially when the number or quality of the measurements is low, such a change can drastically improve the reconstruction quality. In this section, we will introduce these sparsity-based image reconstruction techniques.

4.1 Sparsity and Compressive Sensing

As we saw in the previous section, recovering an image of $N$ pixels requires on the order of $N$ measurements. The motivating idea behind compressive sensing is that many images can be compressed by applying a suitable linear transform and throwing away small values. Put another way, when expressed in the correct basis, many images have few nonzero coefficients. Mathematically, we can write that

$$\|Lf\|_0 = K \ll N,$$  \hspace{1cm} (4.1)

where $\|f\|_0$ is the number of nonzero elements of $f$. It turns out that having this additional information about $f$ allows us to recover it from a number of measurements much smaller than its number of pixels. Popular choices for the sparsifying transform include finite differences, the Fourier transform, the
discrete cosine transform, and the wavelet transform; we compare a few of these in Figure 4.1.

| Method       | (a) Retaining 13,000 largest transform coefficients | (b) Retaining 2,600 largest transform coefficients | (c) Retaining 1,300 largest transform coefficients |
|--------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|
| DFT          | DCT, 26.9 dB                                     | DCT, 20.6 dB                                     | DCT, 18.8 dB                                     |
| DCT          | Haar, 28.8 dB                                    | Haar, 23.8 dB                                    | Haar, 22.3 dB                                    |
| Haar         | db4, 30.6 dB                                     | db4, 23.1 dB                                     | db4, 20.9 dB                                     |
| db4          |                                                  |                                                  |                                                  |

Figure 4.1: Example of the compressibility of biomedical images. We transform a $512 \times 512$ image using the discrete Fourier transform (DFT), $8 \times 8$ block discrete cosine transform (DCT), Haar wavelet transform (Haar), or Daubechies 4 wavelet transform (db4). We set the smallest 95, 99, and 99.5 percent of the coefficients to zero and reconstruct. With 95 percent compression, the compressed image is indistinguishable from the original. Even with the larger compression ratios, much of the detail of the original is preserved. (Image from The Cancer Imaging Archive (TCIA) [1].)

We begin by considering $L = I$, so that $\|f\|_0 = K$. Given measurements,
\( g = Hf \), we can then formulate the compressive sensing problem,

\[
\arg\min_f \|Hf - g\|^2 \quad \text{s.t.} \quad \|f\|_0 \leq K,
\]

(4.2)

where the \( \ell_0 \) norm, \( \|f\|_0 \), counts the number of nonzero entries of \( f \). In general, (4.2) is very challenging to solve, however under certain conditions on \( H \) \cite{23,24}, the solution is unique and the problem is equivalent to

\[
\arg\min_f \|Hf - g\|^2 + \lambda \|f\|_1,
\]

(4.3)

where \( \|f\|_1 = \sum_i |[f]_i| \), for a particular choice of \( \lambda \). (Note the change from the \( \ell_0 \) to \( \ell_1 \) norm; this is what makes the problem tractable.) If we reintroduce the sparsifying transform, \( L \), we find

\[
\arg\min_f \|Hf - g\|^2 + \lambda \|Lf\|_1.
\]

(4.4)

Note that the while the formulation (4.4) is quite similar to the Tikhonov formulation (3.8), changing to the \( \ell_1 \) norm has important consequences; we contrast these two formulations in Figures 4.2 and 4.3.

Figure 4.2: Comparison of \( \ell_2 \) and \( \ell_1 \) regularization when the number of measurements is not less than the dimension of the reconstruction. The gray point is the unique unregularized solution and the gray ellipse is \( \{ f : \|Hf - g\|^2 \leq \sigma^2 \} \). The solid blue line is a level set of \( \|f\|_2 \), and the blue point marks the \( \ell_2 \) solution of the problem. Likewise, the dashed red line is a level set of \( \|f\|_1 \) and the red point marks the \( \ell_1 \) solution. As expected, the \( \ell_1 \) solution is sparse (because it has only one non-zero element, \( f_1 \)).

The formulation (4.4) is called the analysis form of the regularization, because the matrix \( L \) retrieves the sparse coefficients from the original signal. This is in contrast to the synthesis form,

\[
\arg\min_f \|HL\alpha - g\|^2 + \lambda \|\alpha\|_1,
\]

(4.5)
4.2 Representer Theorems for $\ell_2$ and $\ell_1$ Problems

Another perspective on sparsity-promoting regularization is given by representer theorems, which specify the form of solutions to certain minimization problems. For example, for Tikhonov regularization with $L$ being the identity matrix, we can state the following representer theorem \[26\].

**Theorem 3** (Convex Problem with $\ell_2$ Minimization). The problem

$$\arg\min_f \|f\|^2 \quad s.t. \quad \|Hf - g\|^2 \leq \sigma^2,$$

has a unique solution of the form

$$f^* = H^T a,$$
for a suitable set of coefficients, \( a \in \mathbb{R}^M \).

The useful insight here is that the solution to an \( \ell_2 \)-regularized problem always has the form of a weighted sum of the original measurement vectors, i.e. the columns of \( H \). Moreover, the number of elements in the sum is equal to the number of measurements, \( M \), so there is no reason to expect \( f \) to be sparse, unless these measurement vectors are themselves sparse in some transform domain. When \( L \) is not the identity, we can state a similar theorem with \( H^T \) replaced by \((L^T L)^{-1} H^T\).

Contrast this with the \( \ell_1 \) representer theorem \cite{26},

Theorem 4 (Convex Problem with \( \ell_1 \) Minimization). The set

\[
\mathcal{V} = \arg \min_f \|f\|_1 \quad s.t. \quad \|Hf - g\|_2^2 \leq \sigma^2,
\]

is convex, compact, and has extreme points of the form

\[
f^* = \sum_{k=1}^K a_k \delta[\cdot - n_k],
\]

for a suitable set of coefficients \( a \in \mathbb{R}^K \) and locations \( n \in \{1, 2, \ldots, N\}^K \) with \( K \leq M \).

Looking at the form of \( f^* \), we see that it is sparse: it has a number of nonzero terms smaller than the number of measurements with amplitudes given by \( a \) and locations given by \( n \). One complications is that the solution to the \( \ell_1 \) problem is not, in general, unique (though with additional conditions it is). Thus, the theorem is stated in terms of the extreme points of the solution set, i.e., those solutions that cannot be expressed as linear combinations of other solutions. When \( L \) is not the identity, a similar theorem can be stated where the extreme points of the solution set are built out of a sparse sum of kernels that depend on \( L \).

### 4.3 Statistical View

Again, we can turn to a statistical perspective on sparsity-promoting regularization. Here, we follow the approach of \cite{27}, which places both \( \ell_1 \) and \( \ell_2 \) regularization along a spectrum of sparse regularizers.

Working for the moment in 1D, we model the signal as a stochastic process specified by the innovation model

\[
u = Lf, \quad (4.6)
\]
where the discrete innovation, \( u \), has i.i.d. elements with an infinitely divisible PDF, \( p_U \), and \( L \) is a whitening operator. The family of infinitely divisible PDFs includes the Gaussian distribution, as well as several sparser (heavy-tailed or with more mass at the origin) distributions, e.g., the Laplace, compound-Poisson, and Cauchy.

Combining this signal model with a forward operator \( H \) and a Gaussian noise model, we can express the likelihood

\[
p(f \mid g) = \frac{p(g \mid f)p(f)}{p(g)} \quad (4.7)
\]

\[
= \frac{p(g - Hf)p(f)}{p(g)} \quad (4.8)
\]

\[
\propto \exp \left( -\frac{\|g - Hf\|^2}{2\sigma^2} \right) \prod_{n=1}^{N} p_U([Lf]_n). \quad (4.9)
\]

Taking the negative log and minimizing (i.e., performing MAP estimation), we have

\[
\hat{f} = \arg\min_f \frac{1}{2}\|g - Hf\|^2 + \sigma^2 \sum_n \Phi_U([Lf]_n), \quad (4.10)
\]

where \( \Phi_U(u) = -\log p_U(u) \) is called the potential function.

| Distribution     | \( p_U(u) \) | \( \Phi_U(u) \) |
|------------------|---------------|-----------------|
| Gaussian         | \( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{u^2}{2\sigma^2}} \) | \( \frac{1}{2\sigma^2} u^2 + C \) |
| Laplace          | \( \frac{\lambda}{2} e^{-\lambda|u|} \) | \( \lambda|u| + C \) |
| Student          | \( \frac{1}{B(r, \frac{1}{2})} \left( \frac{1}{u^2+1} \right)^{r+\frac{1}{2}} \) | \( (r + \frac{1}{2}) \log(1 + u^2) + C \) |

Table 4.1: Infinitely divisible distributions and their corresponding potential functions. \( B \) is the beta function.

We give a few examples of infinite divisible distributions and the corresponding potential functions in Table 4.1. In the first line, note that the Gaussian potential corresponds to the squared Euclidean norm (with scaling and plus a constant); thus, \( \ell_2 \) regularization corresponds to MAP estimation of a stochastic process with a Gaussian innovation. This is a generalization of example from Section 3.4 because here we also have the operator \( L \). In the second line, we see that the Laplace potential corresponds to the \( \ell_1 \) norm, which means that \( \ell_1 \) regularization corresponds to MAP estimation of a stochastic process with a Laplace-distributed innovation. This links with the concept of sparsity because the Laplace distribution is sparser than the Gaussian. The Student distribution (third line of Table 4.1) is the sparsest of the three, and,
again, its potential function is closely related to sparsity-promoting regularization. Specifically, it is an upper bound on the log prior,

$$\sum_{n=1}^{N} \log |[f]_n|,$$  \hspace{1cm} (4.11)

which can be used as a relaxation of the $\ell_0$ norm in the compressive sensing problem \[4.2\] without changing the minima. It also arises in other sparse reconstruction algorithms, including iterative reweighting schemes and limits of $p$ norms with $p$ tending toward zero \cite{28}.

\subsection{4.4 Algorithms}

A wide range of algorithms have been developed for solving problem like (4.4). One key concept for understanding them is that of the proximal operator.

**Definition 10.** The proximal operator, $\text{prox}_\Phi : \mathbb{R}^N \rightarrow \mathbb{R}^N$, of a convex function, $\Phi : \mathbb{R}^N \rightarrow \mathbb{R}$, is defined by

$$\text{prox}_\Phi (u) = \arg \min_{f} \frac{1}{2} \|u - f\|^2 + \Phi(f).$$ \hspace{1cm} (4.12)

Intuitively, it finds an $f$ with a low $\Phi(f)$, while being close (proximal) to $u$.

If $\Phi$ acts pixel by pixel, then the optimization problem that the prox solves can be decomposed into $N$ scalar problems, one for each pixel. And, in many cases, these scalar problems can be solved efficiently and in parallel.

The prox is used in a class of algorithms called forward-backward splitting \cite{29} or iterative shrinkage and thresholding (ISTA) \cite{30} that solve problems of the form

$$\arg \min_{f} J_1(f) + J_2(f),$$ \hspace{1cm} (4.13)

with $J_1$ differentiable and $J_2$ having a fast prox. For example, these approaches can solve the sparse reconstruction problem \[4.4\] when $L$ is the identity. They achieve this by alternating a step down the gradient of $J_1$ with an application of the prox of $J_2$, which can be shown to converge. In fact, this closely related to projected gradient descent (discussed in Section 3.5) because when a constraint is expressed as a regularization functional (which takes the value infinity whenever the constraint is violated) the associated prox is just projection onto the constraint region.

When the operator $L$ is involved, the sparse reconstruction problem is typically much more difficult to solve. A useful paradigm is to split the data
term and regularization terms by introducing an auxiliary variable. Using the sparse MAP formulation (4.10) as an example, splitting results in

\[
\tilde{f} = \arg \min_{f,u} \frac{1}{2} \|g - Hf\|^2 + \sigma^2 \sum_n \Phi_U([u]_n) \quad \text{s.t.} \quad u = Lf. \quad (4.14)
\]

This problem is equivalent to the original one, but written in the form required by the alternating direction method of multipliers (ADMM) \[31\]. The ADMM then specifies an iterative algorithm,

\[
f^{(k+1)} = \arg \min_f \frac{1}{2} \|g - Hf^{(k)}\|^2 + \frac{\rho}{2} \|Lf - u^{(k)}\|^2 \quad (4.15)
\]

\[
u^{(k+1)} = \arg \min_u \sigma^2 \sum_n \Phi_U([u]_n) + \frac{\rho}{2} \|Lf^{(k)} - u\|^2 \quad (4.16)
\]

\[
\alpha^{(k+1)} = \alpha^{(k)} + Lf^{(k)} - u^{(k)}. \quad (4.17)
\]

It turns out that these three problems are significantly simpler to solve than the original. The update (4.15) is a quadratic problem that can be solved using the techniques from classical image reconstruction (Section 3.5), e.g. the CG algorithm. For the update (4.16) we can make use of the prox of \( \Phi_U \), which is known in closed-form for many potential functions. Finally, the update (4.17) only requires application of \( L \).

### 4.5 Summary

Modern reconstruction techniques focus on sparsity-promoting regularization, with the goal of reconstructing images from very few measurements. This is possible only when the image to be recovered can be sparsely represented in some transform domain. See Figure 4.4 for a comparison between classical variational reconstructions and modern sparsity-based ones. The algorithms to solve sparse reconstruction problems often involve splitting the objective function; each iteration of these algorithms requires solving a classical reconstruction problem, which makes them computationally heavy.

#### 4.5.1 Further Reading

Our discussion of regularization is a simplified version of a much more complex story; see [33] for more historical details and mathematical depth. See [34] for an introduction to compressive sensing. For examples of the statistical view presented here applied to real problems, see [27].
Figure 4.4: Comparison of classical regularization and sparsity-promoting regularization (left and right sides of each panel of (d) and (e), respectively). A discrete test image \[32\] (a), is degraded with blur (b), subsampling, and noise (c). Qualitatively, classical regularization fills gaps by smoothing, while sparsity-promoting regularization gives a piecewise constant solution. For this phantom, the latter is superior.
Chapter 5

The Learning (R)Evolution

The field of machine learning, which deals with creating computer programs that can improve with training \[35\], has existed for decades and has even been used for solving image reconstruction problems as early as in the 1980s \[36\]. Beginning with the emergence of one particular machine learning algorithm, the convolutional neural network (CNN), as a powerful and practical tool for object recognition in 2012 \[37\], there has been a surge in interest in applying learning-based methods to a wide variety of problems in image processing and computer vision, including to image reconstruction. Our goal in this section is to give a broad overview of these approaches.

So far, we have imaged a setting where we have measurements, \( g \), and an understanding of the measurement device from which we build a forward model, \( H \). Additionally, we may have some knowledge of the image we want to reconstruct, allowing us to design a regularization term. In the learning formulation, we assume that we also have access to training data: \( T \) pairs of measurements and their corresponding ground truth reconstructions, \( \{g_t, f_t\}_{t=1}^T \). The inclusion of training data changes the task of the engineer from designing a reconstruction procedure, \( R : g \to f \), to designing a learning procedure, \( L : \{g_t, f_t\}_{t=1}^T \to R \).

We have organized the section along a spectrum of how much is learned. On one end of the spectrum are methods where the training data is used to improve a part of a variational method, usually the forward model or the regularization term. On the other end are pipeline or end-to-end methods, where most or all of the reconstruction procedure is a generic regression function with its parameters learned during training. We also discuss a few methods that do not fit neatly into our learning formulation and the problem of how to acquire training data.
5.1 Learning the Forward Model

Instead of relying on a physical model of the imaging setup alone to design the forward model, $H$, we can use training data to estimate the forward model. This approach is advantageous because the accuracy of the forward model has a large impact on the quality of the reconstruction. Learning the forward model is a daunting task for large problems because the number of elements of the matrix is the number of image pixels times the number of measurements—roughly the number of image pixels squared, easily terabytes of data. Often, knowledge of the structure of the forward model can greatly reduce the number of measurements needed to estimate it.

One example [38] comes from deconvolution of microscopy images. Using our knowledge of the underlying physics, we can assume that the system matrix is a convolution, meaning that is a circulant matrix and therefore completely specified by one of its rows, known in microscopy as the point spread function (PSF). By imaging small beads, we can create a training set with images of the form $f_t = [0 \ldots 0 1 0 \ldots 0]^T$. The resulting measurements, $\{g^t\}$, each provide an independent estimate of the PSF that can be averaged to reduce noise.

Another example [39], comes from PET where measurements of a carefully positioned point source was used to estimate the rows of the forward model. While estimating the entire forward model would require approximately four million measurements, the authors were able to use the geometrical symmetries of the system plus a physical model to interpolate the forward model from just 1,599 measurements.

5.2 Learning the Regularization Term

There are a variety of ways to use training data to improve the regularization term. One straightforward approach is to use training to optimally adjust the regularization strength, i.e. tune the parameter $\lambda$. In the approach of [40], the training takes the form of a bilevel optimization,

$$\arg\min_\lambda \sum_{t=1}^T \|R_\lambda(g^t) - f^t\|,$$  \hspace{1cm} (5.1)

with

$$R_\lambda(g) = \arg\min_f \|g - Hf\| + \lambda \|Lf\|_p^p,$$  \hspace{1cm} (5.2)

where $p$ is one or two. The solution of this problem is the value of $\lambda$ that gives optimal performance on the training set. Reconstructions of new images should then use this optimal $\lambda$. The same method can be extended multiple
5.2. LEARNING THE REGULARIZATION TERM

$L$ operators, each with their own weight; e.g. [40] uses the set of $5 \times 5$ discrete cosine transform (DCT) filters.

Going a step further, we can try to learn the form of the potential function, i.e., the $\Phi$ in a regularization term $\Phi(Lf)$. Recalling the ADMM algorithm from Section 4.4, we know that solving a regularized inverse problem involves applying the proximal operator associated with the potential function. We can implicitly specify the potential function by learning the proximal operator, e.g., by parameterizing it using 1D B-splines [41,42]. We can see prox learning as a generalization of tuning the regularization weight: instead of scaling a known function, we deform a parametric function. If the learned functions satisfy mild constraints, the resulting algorithm is guaranteed to converge.

Or, as a different approach, we can learn the filters in the regularization term leading to a learning procedure [43],

$$\arg\min_{L \in C} \sum_{t=1}^{T} \|Lf_t\|_1,$$

(5.3)

where the matrix $L$ must be constrained to lie in a set $C$ in order to avoid the trivial solution $L = 0$. A variation of this approach is to allow for noisy training data, leading to

$$\arg\min_{L \in C, \{h_t\}_t} \sum_{t=1}^{T} \|Lh_t\|_1 \quad \text{s.t.} \quad \|h_t - f_t\| \leq \sigma.$$  

(5.4)

Recalling the connection between the $\ell_1$ norm and sparsity, we can view this learning procedure as seeking a sparsifying transform for the images in the training set. Intuitively, this learned transform should allow for sparser representations (and therefore better regularization) than a fixed one.

We can also learn a sparsifying transform in the synthesis formation. Doing this is known as sparse dictionary learning (or sparse coding) [44]. In order to learn the $L$ in a synthesis formulation (4.5), we solve

$$\arg\min_{L, \{\alpha_t\}_t} \sum_{t=1}^{T} \|L\alpha_t - f_t\| + \|\alpha\|^p_p,$$

(5.5)

where $p$ is zero or one. Although for simplicity we have written the data fidelity term at the whole-image level, it is usually evaluated at the patch level, so that $L$ is a dictionary that can sparsely represent, e.g., $15 \times 15$ images patches. These patches must then be combined, usually by averaging, to create an image. Convolutional sparse coding, which more-closely matches the formulation (5.5) has also been explored [45], as has a patch-based method with additional image-level regularization to smooth patch boundaries [46]. See [47] for an example of this approach used in X-ray CT.
Finally, we can learn the entire regularization term. Again, this can be done using a bilevel optimization \[48\]

\[
\text{arg min}_{\theta} \sum_{t=1}^{T} \| \mathcal{R}_\theta(g_t) - f_t \|,
\]

with \( \mathcal{R}_\theta(g) = \arg \min \| g - Hf \| + \phi(\mathbf{L}_\theta f) \), \( \text{(5.7)} \)

where \( \theta \) is a vector of parameters that controls both the penalty function \( \phi \) and the analysis filters, \( \mathbf{L} \). Another approach \[49, 50\], which is closer in spirit to a constrained minimization problem, is to learn a image-to-image regression function, \( \mathcal{R}_\theta \), typically a CNN, that can mimic a projector onto the set of plausible images, \( \mathcal{S} \), with plausible here being defined by the training data. Though learning such a function may be difficult, it allows us to solve the very attractive

\[
\text{arg min}_{f \in \mathcal{S}} \| Hf - g \|. \quad \text{(5.8)}
\]

A third approach \[51\] is to combine the ideas of sparsity and projection by a learning parametric image (or patch) encoder and decoder pair according to

\[
\text{arg min}_{\theta} \sum_{t=1}^{T} \| f_t - \mathcal{D}_\theta(\mathcal{E}_\theta(f_t)) \| \quad \text{s.t.} \quad \| \mathcal{E}_\theta(f) \|_0 \leq K. \quad \text{(5.9)}
\]

Reconstruction can then be achieved via

\[
\text{arg min}_{f} \| Hf - g \| + \lambda \| f - \mathcal{D}_\theta(\mathcal{E}_\theta(f)) \| \quad \text{s.t.} \quad \| \mathcal{E}_\theta(f) \|_0 \leq K. \quad \text{(5.10)}
\]

### 5.3 Going Beyond the Variational Framework

One alternative to the variational formulation that has been intensely explored in the last few years is to perform a linear reconstruction followed by a learned image-to-image regressor, typically a CNN. The linear reconstructions often result in heavy artifacts when the number of measurements is low (streaks in CT or aliasing in MRI); the job of the regressor is to remove these artifacts; therefore the method is sometimes referred to as learned denoising or artifact removal. Learning takes the form

\[
\text{arg min}_{\theta} \sum_{t=1}^{T} \| \mathcal{R}_\theta(H^\dagger g_t) - f_t \|, \quad \text{(5.11)}
\]

where \( H^\dagger \) is a matrix that is an approximate inverse of the forward model, e.g. back projection, \( H^T \), or filtered back projection, \( WH^T \) or \( H^T W \), for some
appropriate matrix $W$. The same idea can be used with mixtures of linear (or nonlinear) reconstructions, with the idea that the regressor will learn to spatially vary the regularization strength based on the image content\[52]. The main advantage of these methods, as compared to fully end-to-end methods, is that the design of the regressor is simplified: because it always acts on images (rather than vectors of measurements), it can employ well-studied image processing tools such as multiscale processing and convolutions. This approach has been implemented with a variety of network architectures for X-ray CT\[53,56\] and MRI\[57\] reconstruction.

A common extension to the learned denoising approach is to also learn a data fidelity term. This can be done via perceptual loss\[58\], where images are used as inputs to a pretrained CNN and the distance between them is computed as the Euclidean distance between their intermediate feature vectors; the intuition being that a network trained, e.g., for object recognition, should building perceptually meaningful intermediate representations of the images it operates on. Or, adversarial loss can be employed. In this case, an image-to-scalar regressor, called a discriminator, is trained to give low values to real (in the training set) images and high values to fake (reconstructed images). The discriminator is trained alongside the reconstruction algorithm, pushing it to create ever more plausible reconstructions. See\[59–61\] for examples of these approaches for biomedical image reconstruction.

Finally, we can consider learning the entire reconstruction pipeline end-to-end,

$$\arg\min_\theta \sum_{t=1}^T \| R_\theta(g_t) - f_t \|.$$  \hspace{1cm} (5.12)

Here, all of the work goes into designing the structure of $R_\theta$. One approach is to be problem agnostic, e.g., using a fully-connected neural network\[62\] or a combination of a fully-connected neural network and a CNN\[63\]. An alternative design strategy, called unrolling, is to build the network structure from a preexisting reconstruction algorithm by turning its iterations into layers in the network. In effect, the network is a traditional algorithm for solving an inverse problem, except that every linear operation is learned (rather than being specified by $H$) and the number of iterations/layers is usually low, less than ten, to prevent the number of parameters in the network from becoming too large. For example,\[53\] notes that when $H^2 H$ and $L$ are both convolutions, the typical iterate solutions to the $\ell_1$-regularized variational problem have the form of a CNN, this equivalence allows a some learned denoising algorithms to be interpreted as unrolled iterative methods. Reference\[64\] provides another perspective, proposing to also learn the nonlinear activation functions inside the neural network as a way to learn the regularization functionals, and\[65\] explores the idea of parameterizing the ADMM algorithm. An intermediate approach is to allow $H$ to remain fixed, but to learn, e.g.,
the gradient step in unrolled gradient descent \cite{66} or the proxs in an unrolled primal-dual algorithm \cite{67}.

The main advantage of learning most or all of the reconstruction pipeline is that the results can be excellent, outperforming a well-tuned $\ell_1$ result both quantitatively and qualitatively. The algorithms are also extremely fast compared to iterative methods, with subsecond reconstruction times. The drawback is that training the algorithms is laborious: training can take days and is not guaranteed to converge, meaning that it must be repeated to find suitable hyperparameters (including training parameters as well as the network architecture itself). These problems are compounded by the fact that the network may need to be trained again if aspects of the imaging scenario, e.g., noise level or body part, change \cite{68}. Another drawback is that as more of the reconstruction pipeline is learned, the data term seems to disappear more and more, meaning that the returned solution may not be consistent with the measurements. On one hand, this mismatch may be advantageous: if the forward model is inaccurate, enforcing data consistency may make solutions worse (farther from the ground truth). On the other hand, reconstructions that are plausible-looking without explaining the measurements are hardly useful images. One way of enforcing consistency is to mix the learned reconstruction with a conventional one \cite{69}

$$
\arg\min_f \|g - Hf\| + \lambda\|f - R(g)\|. 
$$

(5.13)

5.4 Other Designs

There are many more designs that do not fit well into the previous sections. We share a few here to give a sense of the diversity of the field.

In many problems, we might be able to access high-quality measurements, $g_{HQ}$, at the cost of a longer or more-costly acquisition, e.g., by taking longer X-ray exposures or sampling more of the $k$-space in MRI. The goal is to use these measurements to learn to reconstruct from low-quality (low dose, fast) measurements. This situation does not fit with our formulation because we assume that the training measurements are the same as the measurements we aim to reconstruct from. One solution is simply to generate a training pair $(f_t, g_{HQ})$ by performing a reconstruction (usually linear) on $g_{HQ}$ and by downsampling $g_{HQ}$. But, $g_{HQ}$ can also be used explicitly to learn how to inpaint missing measurements \cite{70}

$$
\arg\min_\theta \sum_{t=1}^T \|R_\theta(D(g_{HQ}^t)) - g_{HQ}^t\|, 
$$

(5.14)

where $D$ is a model of measurement degradation (e.g., downsampling) and $R_\theta$ is a parametric inpainting function. Then, to reconstruct from low-quality
measurements, we apply $R_\theta$ to generate high-quality measurements and reconstruct using a conventional method. Another, similar, variation on the setting is learning to regress from a reconstruction from one measurement type to a reconstruction from another, for example, to infer one MRI scan type from another [71,72].

Another closely-linked topic that does not exactly fit the paradigm is blind or semiblind image reconstruction. These involve forward model learning without training data. Although the problem is superficially impossible—recover $f$ from $g = Hf + n$ with unknown $H$—it can be solved provided enough constraints on $H$ and prior knowledge about $f$. The archetypal example is blind deconvolution [73], where, as with PSF estimation from training, the fact that $H$ is a convolution is key. Many learning algorithms can be adapted to work without training in a similar way, by alternating a reconstruction step with a learning step that uses the reconstructed images themselves as training; the challenge is that such a procedure can easily diverge and give nonsense results.

5.5 Where to Get the Training Data

A key ingredient in any learning method is the training data. But, how can we acquire training data for inverse problems without already having a working reconstruction algorithm to make images from measurements? If we have such a system, why do we need to design another one? The easiest route out of this chicken-and-egg problem, as we already mentioned in the previous section, is to work in the low-dose (sometimes called compressed-sensing) regime, where we can access high-quality measurements (and reconstructions) at the cost of longer scans. We can then simulate low-quality measurements by degrading the high-quality ones. Thus we train a low-dose reconstruction system from a high-dose one.

But, this is not the only way. In [74], the authors acquire training data by using both a low and high numerical aperture microscope, thus removing the need to simulate low-quality measurements. The cost of this more-realistic training data is that the images from both microscopes needs to be correctly registered before learning. Or, training data can come from a highly-accurate physical simulation [71] that may be impossible to use directly as a forward model because of its computational cost or nonlinearity.

5.6 Summary

Learning-based reconstruction involves using a training set to tune a parameteric reconstruction algorithm. There is a huge variety in the specific ar-
chitecture of these algorithms, with some being closely related to direct or variational methods and others not. In any case, creating a suitable training set is a key challenge.

5.6.1 Further Reading

The recent special issue [75] focuses on original research on learning-based tomographic reconstruction. Reference [76] provides a review of CNNs in the context of biomedical imaging, and [77] provides another perspective with a focus on general image processing. For a long-term perspective on learning in imaging, see [78], and for a look at how these learning approaches might affect the practice of radiology, see [79].
Chapter 6

Conclusion

In this tutorial, we have given a roughly chronological overview of biomedical image reconstruction. We began with a toolbox of mathematical operators that can be used to build models of many physical imaging systems. We then showed how to use these forward models to solve reconstruction problems, either via direct inversion or variational formulations, with a focus on the paradigm of sparsity-promoting regularization. We ended by covering some of the many ways that training data can be used to develop new reconstruction methods via machine learning. Here, we provide a brief comparison of these methods, commenting on the strength of their theoretical underpinnings, their complexity, speed, performance, and robustness. We conclude with comments on future directions for the field.

6.1 Comparisons

Our theoretical understanding is strongest for classical reconstruction methods and weakest for recent learning-based methods. The theory behind both classical reconstruction algorithms and many sparsity-based ones is, for the most part, settled. Thus, we can often make definitive statements about these algorithms, e.g., we know that certain iterative procedures converge to global optimums. We can sometimes even state that a method is optimal for a certain class of problems. That said, there is plenty of work to be done, especially in pushing our understanding to ever more complex and realistic data models. There has been significant progress in explaining the strong performance of learning-based methods on many imaging processing tasks, including work from the perspective of approximation theory [80], unrolling [81], and invariants [82]. Despite these efforts, it seems that theory will continue to lag behind practice for years to come. We should accept the possibility that learning ap-
proaches may never be understood as fully as we might like, and, further, that understanding may become increasingly secondary to performance. The fact that the best-understood methods are not the top performers is not a new state of affairs in imaging, but it is understandably disturbing to see the gap widen; see 83 for a clear expression of the sentiment.

The trend in complexity among the various reconstruction methods is the same: classical methods are simple to implement and tune, while learning-based methods are notoriously hard to train properly. For classical methods, the number of parameters is small and can sometimes be calibrated, e.g., by estimating the measurement noise. Superficially, learning can remove the problem of parameter selection by fixing parameters during training. However, training requires its own set of parameters (sometimes called hyperparameters): the initialization, gradient step size, etc. Because the training problem is usually nonlinear and nonconvex, all of these parameters affect the result. However, in a practical settings, such as industrial or medical imaging, it can be worth investing the engineering effort to train the system.

In terms of speed, conventionally we think of trading speed for reconstruction quality: direct reconstruction is fast but gives the worst results, increasingly sophisticated regularization improves these results at the cost of increased runtime. Learning-based methods upend this order. All of the runtime is spent during training, which happens only once, after which reconstructions are nearly instant. In fact, this happens because only network architectures that allow very fast evaluations can be trained at all. It is this mode of thinking—fix the computational budget, then optimize the algorithm within it—that underlies the unrolling concept 81, and it is a valuable one.

In terms of reconstruction performance, direct methods are excellent when the number of measurements is high and the noise is low; regularization and learning only make sense in the low-dose (high-noise) regime. (In the full-dose regime, it is not clear how learning can be used at all because the training data itself would need to come from some reconstruction algorithm.) In the low-dose regime, however, regularization or learning is key, and learning-based methods show the best performance on almost every reconstruction benchmark. One perspective is that (when sufficient training is available) learning-based methods provide an upper bound on performance, which variational methods might achieve if their regularization and data terms are perfectly suited to the task at hand. One example of this is in 53, where the variational approach beats the learned one when the data fits the regularization model (total variation (TV)-based reconstruction of piecewise constant images).

In terms of robustness, learning-based algorithms are generally the weakest. As we discussed in the last section, slight changes in the imaging parameters (noise level, body part, etc.) can severely reduce the performance of
the network, or, worse, lead to the appearance of realistic-looking structures in the image that are not supported by the measurements at all. Also, in the context of classification, there is the concept of adversarial examples \cite{84}, which are images that are designed to fool a neural network. While there is debate about why adversarial examples can be created and what their existence means, they might indicate that a learning-based image reconstruction algorithm will occasionally (and inexplicably), give a spurious result. For direct and variational methods, on the other hand, the reconstruction quality is consistent across datasets and degrades smoothly as the measurements degrade. The degradation has a predictable form: noise or specific artifacts (e.g., aliasing in MRI, streaks in CT).

6.2 Future Directions

The future of the field remains to produce the highest-quality image using a fixed imaging budget, e.g., scanning time or radiation dose. Doing this will continue to require realistic physical models, though these models may play a new role in reconstruction. For example, they might be used to create training data, or they themselves might be learned from training data. The focus on signal modeling will remain, and should be pursued both with sophisticated mathematical modeling and with data-driven approaches. Finally, we will always seek practical algorithms with few hand-tuned parameters.

Specifically in the context of biomedical imaging, we advocate for a shift towards task-oriented evaluations. The images that we reconstruct in the biomedical domain are usually used for a specific, predefined purpose, e.g., to grade a tumor or to measure the quantify the effect of a genetic manipulation on a model organism. In this setting, the traditional measures of reconstruction performance (signal-to-noise ratio (SNR), mean squared error (MSE), structural similarity index (SSIM)) are insufficient and should be replaced with task-oriented evaluations. Such measures may be expensive (if they require human experts) and they may reveal difficult truths (e.g., that even a few dB improvement in SNR is meaningless), but they will be critical in creating algorithms that truly advance the state of biomedical imaging in practice.

Ultimately, future reconstruction algorithms should employ educated learning: a careful fusion between model-based approaches and data-driven ones. We believe that this approach will provide both excellent performance and the robustness and performance guarantees needed for biomedical applications. In biology, this will enable high-quality imaging of a wider range of biological structures and processes; in medicine, this will allow diagnostically useful images to be created more quickly, cheaply, and with less harm to the patient.
Bibliography

[1] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, “The cancer imaging archive (TCIA): Maintaining and operating a public information repository,” *Journal of Digital Imaging*, vol. 26, pp. 1045–1057, July 2013.

[2] D. A. Gutman, M. Khalilia, S. Lee, M. Nalisnik, Z. Mullen, J. Beezley, D. R. Chittajallu, D. Manthey, and L. A. Cooper, “The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research,” *Cancer Research*, vol. 77, pp. e75–e78, Oct. 2017.

[3] L. Mercier, R. F. D. Maestro, K. Petrecca, D. Araujo, C. Haegelen, and D. L. Collins, “Online database of clinical MR and ultrasound images of brain tumors,” *Medical Physics*, vol. 39, pp. 3253–3261, May 2012.

[4] S.-W. Lee and R. Mittra, “Fourier transform of a polygonal shape function and its application in electromagnetics,” *IEEE Transactions on Antennas and Propagation*, vol. 31, pp. 99–103, Jan. 1983.

[5] A. Webb, *Introduction to Biomedical Imaging*. Hoboken, NJ: Wiley-IEEE Press, 1 edition ed., Dec. 2002.

[6] J. A. Fessler, “Model-Based Image Reconstruction for MRI,” *IEEE Signal Processing Magazine*, vol. 27, pp. 81–89, July 2010.

[7] M. Davidson, “Microscopyu,” 2001.

[8] A. Diaspro and M. A. M. J. van Zandvoort, *Super-Resolution Imaging in Biomedicine*. CRC PR INC, 2016.

[9] A. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*. Classics in Applied Mathematics, New York, NY: Society for Industrial and Applied Mathematics, 2001.
[10] D. L. Bailey, D. W. Townsend, P. E. Valk, and M. N. Maisey, eds., Positron Emission Tomography: Basic Sciences. New York: Springer, 2003 edition ed., Apr. 2005.

[11] J. Frank, ed., Electron Tomography. Springer New York, 2006.

[12] E. Wolf, “Three-dimensional structure determination of semi-transparent objects from holographic data,” Optics Communications, vol. 1, pp. 153–156, Sept. 1969.

[13] A. V. Oppenheim and R. W. Schafer, Discrete-Time Signal Processing. ADDISON WESLEY PUB CO INC, 2009.

[14] M. Vetterli, J. Kovacevic, and V. Goyal, Foundations of Signal Processing. Cambridge, United Kingdom: Cambridge University Press, Oct. 2014.

[15] M. Unser, E. Soubies, F. Soulez, M. McCann, and L. Donati, “Global-BioIm: A Unifying Computational Framework for Solving Inverse Problems,” in Imaging and Applied Optics 2017, (San Francisco, California), p. CTu1B.1, June 2017.

[16] G. N. Ramachandran and A. V. Lakshminarayanan, “Three-dimensional reconstruction from radiographs and electron micrographs: Application of convolutions instead of fourier transforms,” Proceedings of the National Academy of Sciences, vol. 68, pp. 2236–2240, Sept. 1971.

[17] R. Ciak, B. Shafei, and G. Steidl, “Homogeneous penalizers and constraints in convex image restoration,” Journal of Mathematical Imaging and Vision, vol. 47, pp. 210–230, Oct. 2012.

[18] M. Nikolova, “Model distortions in bayesian MAP reconstruction,” Inverse Problems and Imaging, vol. 1, pp. 399–422, Apr. 2007.

[19] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, 2004.

[20] J. Shewchuk, “An introduction to the conjugate gradient method without the agonizing pain,” tech. rep., Carnegie Mellon University, 1994.

[21] C. Vonesch, L. Wang, Y. Shkolnisky, and A. Singer, “Fast wavelet-based single-particle reconstruction in cryo-EM,” in 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, (Chicago, Illinois), Mar. 2011.

[22] M. T. McCann, M. Nilchian, M. Stampanoni, and M. Unser, “Fast 3d reconstruction method for differential phase contrast X-ray CT,” Optics Express, vol. 24, p. 14564, June 2016.
[23] D. L. Donoho, “Compressed sensing,” *IEEE Transactions on Information Theory*, vol. 52, pp. 1289–1306, Apr. 2006.

[24] E. Candès, J. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Comm. Pure Appl. Math.*, vol. 59, pp. 1207–1223, Aug. 2006.

[25] M. Elad, P. Milanfar, and R. Rubinstein, “Analysis versus synthesis in signal priors,” *Inverse Problems*, vol. 23, pp. 947–968, Apr. 2007.

[26] M. Unser, J. Fageot, and H. Gupta, “Representer theorems for sparsity-promoting $\ell_1$ regularization,” *IEEE Transactions on Information Theory*, vol. 62, pp. 5167–5180, Sept. 2016.

[27] M. Unser and P. D. Tafti, *An Introduction to Sparse Stochastic Processes*. Cambridge; New York: Cambridge University Press, 1 edition ed., Oct. 2014.

[28] D. Wipf and S. Nagarajan, “Iterative reweighted $\ell_1$ and $\ell_2$ methods for finding sparse solutions,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 317–329, Apr. 2010.

[29] P. L. Combettes and V. R. Wajs, “Signal recovery by proximal forward-backward splitting,” *Multiscale Modeling & Simulation*, vol. 4, pp. 1168–1200, Jan. 2005.

[30] A. Beck and M. Teboulle, “A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems,” *SIAM Journal on Imaging Sciences*, vol. 2, pp. 183–202, Jan. 2009.

[31] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends in Machine Learning*, vol. 3, pp. 1–122, Jan. 2011.

[32] L. A. Shepp and Y. Vardi, “Maximum Likelihood Reconstruction for Emission Tomography,” *IEEE Transactions on Medical Imaging*, vol. 1, pp. 113–122, Oct. 1982.

[33] M. Benning and M. Burger, “Modern regularization methods for inverse problems,” *Acta Numerica*, vol. 27, pp. 1–111, May 2018.

[34] E. Candès, N. Braun, and M. Wakin, “sparse signal and image recovery from compressive samples,” in *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, IEEE, 2007.
[35] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill Education, 1 edition ed., Mar. 1997.

[36] Y. T. Zhou, R. Chellappa, A. Vaid, and B. K. Jenkins, “Image restoration using a neural network,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 1141–1151, July 1988.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision*, vol. 115, pp. 211–252, Dec. 2015.

[38] P. J. Shaw and D. J. Rawlins, “The point-spread function of a confocal microscope: its measurement and use in deconvolution of 3-d data,” *Journal of Microscopy*, vol. 163, pp. 151–165, Aug. 1991.

[39] V. Panin, F. Kehren, C. Michel, and M. Casey, “Fully 3-d PET reconstruction with system matrix derived from point source measurements,” *IEEE Transactions on Medical Imaging*, vol. 25, pp. 907–921, July 2006.

[40] K. Kunisch and T. Pock, “A bilevel optimization approach for parameter learning in variational models,” *SIAM Journal on Imaging Sciences*, vol. 6, pp. 938–983, Jan. 2013.

[41] U. S. Kamilov and H. Mansour, “Learning optimal nonlinearities for iterative thresholding algorithms,” *IEEE Signal Processing Letters*, vol. 23, pp. 747–751, May 2016.

[42] H. Q. Nguyen, E. Bostan, and M. Unser, “Learning convex regularizers for optimal bayesian denoising,” *IEEE Transactions on Signal Processing*, vol. 66, pp. 1093–1105, Feb. 2018.

[43] M. Yaghoobi, S. Nam, R. Gribonval, and M. E. Davies, “Constrained overcomplete analysis operator learning for cosparse signal modelling,” *IEEE Transactions on Signal Processing*, vol. 61, pp. 2341–2355, May 2013.

[44] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Transactions on Image Processing*, vol. 15, pp. 3736–3745, Dec. 2006.

[45] H. Bristow, A. Eriksson, and S. Lucey, “Fast convolutional sparse coding,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, (Portland, OR), IEEE, June 2013.
[46] S. Soltani, M. S. Andersen, and P. C. Hansen, “Tomographic image reconstruction using training images,” *Journal of Computational and Applied Mathematics*, vol. 313, pp. 243–258, Mar. 2017.

[47] Q. Xu, H. Yu, X. Mou, L. Zhang, J. Hsieh, and G. Wang, “Low-dose X-ray CT reconstruction via dictionary learning,” *IEEE Transactions on Medical Imaging*, vol. 31, pp. 1682–1697, Sept. 2012.

[48] Y. Chen, R. Ranftl, and T. Pock, “Insights into analysis operator learning: From patch-based sparse models to higher order MRFs,” *IEEE Transactions on Image Processing*, vol. 23, pp. 1060–1072, Mar. 2014.

[49] J. H. R. Chang, C.-L. Li, B. Poczos, and B. V. K. V. Kumar, “One network to solve them all — solving linear inverse problems using deep projection models,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, (Venice, Italy), IEEE, Oct. 2017.

[50] H. Gupta, K. H. Jin, H. Q. Nguyen, M. T. McCann, and M. Unser, “CNN-based projected gradient descent for consistent CT image reconstruction,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1440–1453, June 2018.

[51] D. Wu, K. Kim, G. E. Fakhri, and Q. Li, “Iterative low-dose CT reconstruction with priors trained by artificial neural network,” *IEEE Transactions on Medical Imaging*, vol. 36, pp. 2479–2486, Dec. 2017.

[52] D. Boublil, M. Elad, J. Shtok, and M. Zibulevsky, “Spatially-Adaptive Reconstruction in Computed Tomography Using Neural Networks,” *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1474–1485, July 2015.

[53] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, “Deep Convolutional Neural Network for Inverse Problems in Imaging,” *IEEE Transactions on Image Processing*, vol. 26, pp. 4509-4522, Sept. 2017.

[54] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang, “Low-dose CT with a residual encoder-decoder convolutional neural network,” *IEEE Trans. Med. Imag.*, vol. 36, pp. 2524–2535, Dec. 2017.

[55] E. Kang, W. Chang, J. Yoo, and J. C. Ye, “Deep convolutional framelet denoising for low-dose CT via wavelet residual network,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1358–1369, June 2018.

[56] Y. Liu and Y. Zhang, “Low-dose CT restoration via stacked sparse denoising autoencoders,” *Neurocomputing*, vol. 284, pp. 80–89, Apr. 2018.
[57] C. M. Hyun, H. P. Kim, S. M. Lee, S. Lee, and J. K. Seo, “Deep learning for undersampled MRI reconstruction,” *Physics in Medicine & Biology*, vol. 63, p. 135007, June 2018.

[58] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution,” in *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, (Cham, Switzerland), pp. 694–711, Springer, Oct. 2016.

[59] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, and D. Firmin, “DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1310–1321, June 2018.

[60] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong, “Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1488–1497, June 2018.

[61] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, “Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1348–1357, June 2018.

[62] D. C. Durairaj, M. C. Krishna, and R. Murugesan, “A neural network approach for image reconstruction in electron magnetic resonance tomography,” *Computers in Biology and Medicine*, vol. 37, pp. 1492–1501, Oct. 2007.

[63] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, “Image reconstruction by domain-transform manifold learning,” *Nature*, vol. 555, pp. 487–492, Mar. 2018.

[64] E. Kobler, T. Klatzer, K. Hammernik, and T. Pock, “Variational networks: Connecting variational methods and deep learning,” in *Lecture Notes in Computer Science*, pp. 281–293, Springer International Publishing, 2017.

[65] J. He, Y. Yang, Y. Wang, D. Zeng, Z. Bian, H. Zhang, J. Sun, Z. Xu, and J. Ma, “Optimizing a parameterized plug-and-play ADMM for iterative low-dose CT reconstruction,” *IEEE Transactions on Medical Imaging*, pp. 1–1, 2018.
[66] J. Adler and O. Öktem, “Solving ill-posed inverse problems using iterative deep neural networks,” *Inverse Problems*, vol. 33, p. 124007, Nov. 2017.

[67] J. Adler and O. Öktem, “Learned primal-dual reconstruction,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1322–1332, June 2018.

[68] F. Knoll, K. Hammernik, E. Kobler, T. Pock, M. P. Recht, and D. K. Sodickson, “Assessment of the generalization of learned image reconstruction and the potential for transfer learning,” *Magnetic Resonance in Medicine*, May 2018.

[69] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, “A deep cascade of convolutional neural networks for dynamic MR image reconstruction,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 491–503, Feb. 2018.

[70] T. Eo, Y. Jun, T. Kim, J. Jang, H.-J. Lee, and D. Hwang, “KIKI-net: cross-domain convolutional neural networks for reconstructing under-sampled magnetic resonance images,” *Magnetic Resonance in Medicine*, vol. 80, pp. 2188–2201, Apr. 2018.

[71] C. Cai, C. Wang, Y. Zeng, S. Cai, D. Liang, Y. Wu, Z. Chen, X. Ding, and J. Zhong, “Single-shot t2 mapping using overlapping-echo detachment planar imaging and a deep convolutional neural network,” *Magnetic Resonance in Medicine*, vol. 80, pp. 2202–2214, Apr. 2018.

[72] K. H. Kim, W.-J. Do, and S.-H. Park, “Improving resolution of MR images with an adversarial network incorporating images with different contrast,” *Medical Physics*, vol. 45, pp. 3120–3131, May 2018.

[73] D. Kundur and D. Hatzinakos, “Blind image deconvolution,” *IEEE Signal Processing Magazine*, vol. 13, pp. 43–64, May 1996.

[74] Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan, “Deep learning microscopy,” *Optica*, vol. 4, p. 1437, Nov. 2017.

[75] G. Wang, J. C. Ye, K. Mueller, and J. A. Fessler, “Image reconstruction is a new frontier of machine learning,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1289–1296, June 2018.

[76] M. T. McCann, K. H. Jin, and M. Unser, “Convolutional neural networks for inverse problems in imaging: A review,” *IEEE Signal Process. Mag.*, vol. 34, pp. 85–95, Nov. 2017.
[77] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos, “Using deep neural networks for inverse problems in imaging: Beyond analytical methods,” *IEEE Signal Processing Magazine*, vol. 35, pp. 20–36, Jan. 2018.

[78] G. Wang, “A perspective on deep imaging,” *IEEE Access*, vol. 4, pp. 8914–8924, 2016.

[79] G. Zaharchuk, E. Gong, M. Wintermark, D. Rubin, and C. Langlotz, “Deep learning in neuroradiology,” *American Journal of Neuroradiology*, Feb. 2018.

[80] K. Hornik, “Approximation Capabilities of Multilayer Feedforward Networks,” *Neural Networks*, vol. 4, pp. 251–257, Mar. 1991.

[81] K. Gregor and Y. LeCun, “Learning fast approximations of sparse coding,” in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, (Haifa, Israel), June 2010.

[82] S. Mallat, “Understanding deep convolutional networks,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, p. 20150203, Mar. 2016.

[83] M. Elad, “Deep, Deep Trouble,” *SIAM News*, vol. 50, p. 12, May 2017.

[84] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *International Conference on Learning Representations*, (Banff, Canada), Apr. 2014.