PARSIMONIOUS MODELS FOR INVERSE PROBLEMS

BY

LUKE PFISTER

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Yoram Bresler, Chair
Professor Rohit Bhargava
Assistant Professor Ivan Dokmanić
Professor P. Scott Carney, University of Rochester

Abstract

This dissertation can be coarsely divided into two parts: Chapters 1 and 2 study the problem of the multidimensional filter bank design and data-driven adaptation, while Chapters 3 to 5 focus on variations of optical tomography.

Chapter 1 describes a fast way to estimate the extremal values of a trigonometric polynomial given samples from the polynomial. This work came about from a simple question: *Can we determine whether the Discrete-Time Fourier Transform of a multidimensional discrete index signal reaches zero, given only its Discrete Fourier Transform?* The answer is yes— provided that the signal has small support and its samples do not vary too much. This property unlocks new possibilities for the numerical design of multidimensional, multirate, perfect reconstruction filter banks; we conclude by designing a curvelet-like filter bank.

Chapter 2 focuses on data-adaptive sparse representations; that is, a sparse representation learned directly from the data itself. These representations are usually described as modeling and acting on small image patches. We show that many of the existing sparse representations can instead be thought of as filter banks, thus linking the local properties of a patch-based model to the global properties of a convolutional model. We then use the results on trigonometric polynomials developed in Chapter 1 as the foundation for a new algorithm to learn perfect reconstruction filter banks that sparsify data. Our learned model outperforms local, patch-based transform learning approaches in image denoising tasks while benefiting from additional flexibility in the design process.

Chapter 3 marks the transition to the second family of topics in this dissertation. In this chapter, we review a particular optical tomographic imaging: Interferometric Synthetic Aperture Microscopy (ISAM). ISAM allows for rapid, non-invasive imaging of quasi-transparent objects in three spatial dimensions from measurements of back-scattered light. In this modality, volumetric images are formed by solving the inverse scattering problem using perturbative methods. The resulting image reconstruction algorithms have efficient numerical implementations.

The usual ISAM image reconstruction algorithms are well-suited for data collected from a single focal plane, with Tikhonov regularization, and/or if Gaussian noise is present. In these situations a non-iterative image reconstruction algorithm is applicable. However, when an iterative solution is required, the perturbative ISAM model leads to artifacts in the reconstructed image.

In Chapter 4, we present a new approximation to the ISAM forward model. This model facilitates the combination of fast numerical algorithms and iterative image reconstruction. We construct the singular value decomposition of our new approximate ISAM operator and investigate the resolution of the imaging system.

In Chapter 5, we combine ISAM with imaging spectroscopy to determine spatial morphology and chemical composition in three spatial dimensions. We assume the target has a low-rank structure; physically, this implies the target is composed of a few distinct chemical species. We call this the *N-species* approximation. We use this low-rank structure to reduce the amount of data needed to solve the inverse scattering problem.

To Jennifer.

Acknowledgments

I have had exceptional mentorship throughout my time at the University of Illinois. First and foremost, I must thank my advisor, Yoram Bresler. He taught me how to think like a mathematician, and an engineer; and, importantly, how and when to switch between these roles. Yoram has invested tremendous time and energy into helping me to grow my technical skills as a researcher, and has taught me how to effectively communicate my ideas.

I took Scott Carney's course, "Inverse Problems in Optics", because I was interested in the first half of the title. It was a great decision. Scott sparked my interest in optical inverse problems, and introduced me to the problems that are studied in the second half of this thesis. He has been a wonderful teacher, mentor, and friend.

I received the Andrew T. Yang fellowship in 2014 and 2015. This award funded my transition to optical inverse problems. I am immensely grateful for this opportunity.

While preparing my application for the Andrew T. Yang fellowship, I was introduced to Rohit Bhargava. He welcomed me into his group, and I have learned much from attending his group meetings and collaborating with his students. Rohit broadened my view of the research landscape, introduced me to new research areas, and has been an excellent motivator. I am grateful for the many discussions we have had.

I would like to thank the members of Yoram's research group: Yanjun Li, Bihan Wen, Saiprasad Ravishankar, and Kiryung Lee for many fruitful discussions. I shared an office with Yanjun and Bihan for several years. Our conversations were the most valuable I had during my time in graduate school. I must also give special thanks to Kiryung, who encouraged me at a time when I needed it the most. I must also thank my friends: Jonathan Ligo, with whom I shared a love of coffee and fountain pens, Andrew Bean, who introduced me to rock climbing, as well as Trong Nguyen, Patrick Johnstone, Ben Chidester, and many others. I have learned from each of you.

The Coordinated Science Laboratory has been my home for many years. I have loved my time in the building, thanks in large part to the support staff. Peggy Wells, Brenda Roy, Chris Holt, and Angie Ellis have always treated me with kindness and warmth.

My family is the foundation that has made this work possible. My sister, Danielle, has grown into a wonderful teacher. She strikes the perfect balance of compassion and determination.

I must thank my parents, Mark and Tammy, for instilling in me the value of hard work and determination. They remain my greatest role models.

Finally, my deepest gratitude goes to my wife, Jennifer. She has been by my side for every step of the process, and has been endlessly patient and supportive throughout the grind. I could not have completed this work without her.

Table of Contents

Chapter 1

Bounding Multivariate Trigonometric Polynomials with Applications to Filter Bank Design

1.1 Introduction

1.1.1 Motivation

Trigonometric polynomials are intimately linked to discrete-time signal processing, arising in problems of controls, communications, filter design, and super resolution, among others. For example, the Discrete-Time Fourier Transform (DTFT) converts a sequence of length n into a trigonometric polynomial of degree $n-1$. Multivariate trigonometric polynomials arise in a similar fashion, as the d-dimensional DTFT yields a d-variate trigonometric polynomial.

The extremal values of a trigonometric polynomial are often of interest. In an Orthogonal Frequency Division Multiplexing (OFDM) communication system, the transmitted signal is a univariate trigonometric polynomial, and the maximum modulus of this signal must be accounted for when designing power amplifiers [1]. The maximum modulus of a trigonometric polynomial is related to the stability of a control system in the face of perturbations [2]. The maximum gain and attenuation of a Finite Impulse Response (FIR) filter are the maximum and minimum values of a real and non-negative trigonometric polynomial. Unfortunately, determining the extremal values of a multivariate polynomial given its coefficients is NP-hard [3, 4].

An approximation to the extremal values can be found by discretizing the polynomial and performing a grid search, but this method is sensitive to the discretization level. Instead, one can try to find the extremal values using an optimization-based approach. However, iterative descent algorithms are prone to finding local optima as a generic polynomial is not a convex function. The sum-of-squares machinery provides an alternative approach: extremal values of a polynomial can be found by solving a hierarchy of semidefinite program (SDP) feasibility problems [2, 4, 5]. Truncating the sequence of SDPs provides a lower (or upper) bound to the minimum (or maximum) of the polynomial. However, the size of the SDPs grows exponentially in the number of variables, d, and polynomially in the degree, n, limiting the applicability of this approach.

In many applications we have access to samples of the polynomial rather than to the coefficients of the polynomial itself. Equally spaced samples of a trigonometric polynomial arise, for instance, when computing the Discrete Fourier Transform (DFT) of a sequence. Given enough samples, the polynomial can be evaluated at any point by periodic interpolation, and thus grid search or optimization-based approaches can still be used; however, the previously described issues of discretization error, local minima, and complexity remain.

In this chapter, we derive simple estimates for the extremal values of a multivariate trigonometric polynomial directly from its samples, *i.e.* with no interpolation step. For a complex polynomial we provide an upper bound on its modulus, while for a real trigonometric polynomial we provide upper and lower bounds. Upper bounds of this style have been derived for univariate trigonometric polynomials—our work provides an extension to the multivariate case. We describe two sample applications that benefit from our lower bound and from the extension to multivariate polynomials.

(i) **Design of Perfect Reconstruction Filter Banks**.

A multi-rate filter bank in d dimensions is characterized by its polyphase matrix, $H(z) \in \mathbb{C}^{m \times n}$, where each entry in the matrix is a d-variate Laurent polynomial[1] in $z \in \mathbb{C}^d$ [6].

Many important properties of the filter bank can be inferred from the polyphase matrix. A filter bank is said to be *perfect reconstruction* (PR) if any signal can be recovered, up to scaling and a shift, from its filtered form. The design and characterization of multirate filter banks in one dimension is well understood, but becomes difficult in higher dimensions due to the lack of a spectral factorization theorem [7–11]. The perfect reconstruction condition is equivalent to the strict positivity of the real trigonometric polynomial $p_H(\omega) = \det\left(H^*(e^{j\omega})H(e^{j\omega})\right)$ [6, 12]. The lower bounds developed in this chapter provide a sufficient condition to verify the perfect reconstruction property from samples of $p_H(\omega)$ which are easily obtained using the DFT.

(ii) **Estimating the Smallest Eigenvalue of a Hermitian Block Toeplitz Matrix with Toeplitz Blocks.**

Toeplitz matrices describe shift-invariant phenomena and are found in countless applications. Toeplitz matrices model convolution with a finite impulse response filter, and the covariance matrix formed from a random vector drawn from a wide-sense stationary (WSS) random process

[1] A Laurent polynomial allows negative powers of the argument.

is symmetric and Toeplitz. An $n \times n$ Toeplitz matrix is of the form

$$X_n = \begin{bmatrix} x_0 & x_{-1} & x_{-2} & \cdots & x_{-n+1} \\ x_1 & x_0 & x_{-1} & & \\ x_2 & x_1 & x_0 & & \vdots \\ \vdots & & & \ddots & \\ x_{n-1} & & & \cdots & x_0 \end{bmatrix},$$

and a Hermitian symmetric Toeplitz matrix satisfies $x_i^* = x_{-i}$. Associated with X_n is the trigonometric polynomial [2]

$$\hat{x}(\omega) = \sum_{k=-n}^{n} x_k e^{j\omega k}, \qquad -\pi \leq \omega < \pi,$$

with coefficients

$$x_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{x}(\omega) e^{-jk\omega} dt, \qquad k \in \mathbb{Z}. \tag{1.1}$$

The polynomial $\hat{x}$ is known as the *symbol* of X_n. If the symbol is real then X_n is Hermitian, and if $\hat{x}$ is strictly positive then X_n is positive definite.

A vast array of literature has examined the connections between a real symbol $\hat{x}$ and the eigenvalues of the Hermitian Toeplitz matrices X_n as $n \to \infty$; see [13, 14] and references therein. One result of particular interest states that the eigenvalues of X_n are upper and lower bounded by the supremum and infimum of the symbol.

The smallest eigenvalue of a Toeplitz matrix is of interest in many applications [15–17], and there are several iterative algorithms to efficiently calculate this eigenvalue [18]. We propose a non-iterative estimate of the smallest and largest eigenvalues of X_n by first bounding the eigenvalues in terms of the symbol, then bounding the symbol in terms of the entries of X_n.

Shift invariant phenomena in two dimensions are described by Block Toeplitz matrices with Toeplitz Blocks (BTTB). The symbol for a BTTB matrix is a bi-variate trigonometric polynomial, and the bounds developed in this chapter hold in this case.

1.1.2 Notation

For a set $\mathbb{X}$, let $\mathbb{X}^d$ be the d-fold Cartesian product $\mathbb{X} \times \ldots \times \mathbb{X}$. Let $\mathbb{T} = [0, 2\pi]$ be the torus and $\mathbb{Z}$ be the integers. The set $\{0, \ldots N-1\}$ is written $[N]$. We denote the space of d-variate trigonometric

[2] This differs from the usual approach of describing Toeplitz matrices, wherein a Toeplitz matrix of size n is generated according to (1.1) for an underlying symbol and the behavior as $n \to \infty$ is investigated. Here, we work with a Toeplitz matrix of fixed size.

polynomials with maximum component degree n as

$$T_n^d \triangleq \operatorname{span}\left\{e^{jk\cdot\omega} : \omega \in \mathbb{T}^d, k \in \mathbb{Z}^d, \|k\|_\infty \leq n\right\},$$

where $x \cdot y \triangleq \sum_{i=1}^d x_i y_i$ is the Euclidean inner product and $\|k\|_\infty = \max_{1\leq i\leq d} |k_i|$. An element of T_n^d is explicitly given by

$$p(\omega) = \sum_{k_1=-n}^{n} \cdots \sum_{k_d=-n}^{n} c_{k_1\ldots k_d} e^{jk_1\omega_1} \ldots e^{jk_d\omega_d}.$$

If the coefficients satisfy $c_{k_1,\ldots,k_d} = c^*_{-k_1,\ldots,-k_d}$, then $p(\omega)$ is real for all ω and p is said to be a *real trigonometric polynomial*. We denote the space of real trigonometric polynomials by $\bar{T}_n^d$. For $p \in T_n^d$ let $\|p\|_\infty = \max_{\omega\in\mathbb{T}^d} \left|p(\omega)\right|$. We write the set of N equidistant sampling points on $\mathbb{T}$ as

$$\Theta_N \triangleq \left\{\omega_k = k\frac{2\pi}{N} : k = 0, \ldots, N-1\right\},$$

and on $\mathbb{T}^d$ as Θ_N^d, given by the d-fold Cartesian product $\Theta_N \times \ldots \times \Theta_N$. The maximum modulus of p over Θ_N^d is

$$\|p\|_{N^d,\infty} \triangleq \max_{\omega\in\Theta_N^d} \left|p(\omega)\right|.$$

1.1.3 Problem Statement and Existing Results

Let $p \in \bar{T}_n^d$. Our goal is to find scalars $a \leq b$, depending only on N, d, and the N^d samples $\left\{p(\omega) : \omega \in \Theta_N^d\right\}$, such that

$$a \leq p(\omega) \leq b.$$

For complex trigonometric polynomials, $p \in T_n^d$, we want an upper bound on the modulus; a lower bound on the modulus can be obtained by considering the real trigonometric polynomial $p' \in \bar{T}_n^{2d} : \omega \mapsto \left|p(\omega)\right|^2$.

By the periodic sampling theorem (Lemma 1.4), trigonometric interpolation perfectly recovers $p \in T_n^d$ from $(2n+1)^d$ uniformly spaced samples. A standard result of approximation theory states [19, 20]

$$\|p\|_\infty \leq \|p\|_{(2n+1)^d,\infty} \left(\frac{\pi+4}{\pi} + \frac{2}{\pi}\log(2n+1)\right)^d, \tag{1.2}$$

but this becomes weak as the polynomial degree n or the dimension d of its domain increases. A more stable estimate is obtained by using non-uniformly spaced samples. However, in many applications the sampled polynomial is obtained using the DFT, thus providing uniformly spaced samples.

Our aim is to get stronger estimates by using more (uniformly spaced) samples than are required by the periodic sampling theorem. Upper bounds for *univariate* trigonometric polynomials have been developed using this strategy. Let $p \in T_n$. Given an integer m and $N = 2m > 2n+1$ samples of p, Ehlich and Zeller showed [21]

$$\|p\|_\infty \leq \left(\cos\left(\frac{\pi n}{2m}\right)\right)^{-1} \|p\|_{N,\infty} \tag{1.3}$$

and this bound is sharp if n is a divisor of m.

Wunder and Boche developed a more flexible bound: given $N \geq 2n+1$, they showed [22]

$$\|p\|_\infty \leq \sqrt{\frac{N+2n+1}{N-(2n+1)}}\|p\|_{N,\infty}. \tag{1.4}$$

Zimmermann *et al.* refined this bound to [1]

$$\|p\|_\infty \leq \frac{\|p\|_{N,\infty}}{\sqrt{1-\alpha}}, \tag{1.5}$$

where $\alpha = 2n/N$. The quantity α^{-1} is almost equal to the oversampling factor $\frac{N}{2n+1}$, and plays the same role: α is a decreasing function of N, and for $N \geq 2n+1$, we have $\alpha < 1$.

The bounds (1.2) to (1.5) each have the form:

$$\|p\|_\infty \leq C^d_{N,n}\|p\|_{N^d,\infty}, \tag{1.6}$$

where $C^d_{N,n}$ is a real, non-negative constant that depends on N, n and, in the case of (1.2), d. In the univariate case, Zimmermann *et al.* studied the optimal value of $C_{N,n}$ and showed that it depends only on N/n [1]. They also characterized *extremal* polynomials, for which (1.6) holds with equality, and discussed a Remez-like algorithm to construct such polynomials for given N and n.

1.1.4 Contributions

Our contributions can be summarized as follows: (i) we develop upper bounds of the form (1.6) for *multivariate* trigonometric polynomials; these include both a multivariate extension of the bound (1.5), as well as a tighter bound for the case of low oversampling ($N \approx 2n+1$); (ii) we specialize and strengthen the bounds for real polynomials; and (iii) we derive a lower bound for real trigonometric polynomials.

1.2 Statement of Main Results

In this section we collect our main results; proofs are deferred to Sections 1.3 and 1.4. For simplicity we work with T_n^d, but the results can be easily strengthened by allowing for the component degree to vary in each of the d dimensions.

Our first task is to obtain bounds of the form (1.6) for multivariate trigonometric polynomials. We have a pair of such bounds.

Theorem 1.1. *Let $p \in T_n^d$. Take $N \geq 2n+1$ and set $\alpha = 2n/N$. Then*

$$\|p\|_\infty \leq C_{N,n}^d \|p\|_{N^d,\infty}, \tag{1.7}$$

where

$$\begin{aligned} C_{N,n}^d &\triangleq \frac{\left(\sup\limits_{\omega\in\mathbb{T}}\left\{\sum\limits_{\omega_k\in\Theta_N}\left|\frac{\sin\left(\frac{N\omega}{2}\right)\sin\left(\frac{N-2n}{2}(\omega-\omega_k)\right)}{\sin^2\left((\omega-\omega_k)/2\right)}\right|\right\}\right)^d}{N^d(N-2n)^d} && (1.8)\\ &\leq (1-\alpha)^{-\frac{d}{2}}. && (1.9)\end{aligned}$$

Further,

$$C_{N,n}^d\|p\|_{N^d,\infty} - \|p\|_\infty \leq \left(\frac{dn}{N} + \mathcal{O}((dn/N)^2)\right)\|p\|_\infty.$$

The bound (1.8) involves only a univariate function and can be calculated numerically. Still, the expression is unwieldy; (1.9) is a simpler, but weaker, alternative.

We plot the behavior of $C_{N,n}$, given by (1.8) and (1.9) for the $d=1$ univariate case, in Fig. 1.1. Also shown in Fig. 1.1 are the optimal values of $C_{N,n}$ for integer oversampling factors, given by (1.3), and the values obtained using Zimmermann's Remez-like algorithm [1].

The upper bound (1.7) with $C_{N,n}^d$ given by (1.8) is nearly tight for $N/(2n) < 2$, whereas replacing $C_{N,n}^d$ by its upper bound (1.9) results in a weakening of (1.7) in this regime. This gap makes (1.8) particularly attractive in the d-variate case, where the bounds are raised to the d-th power, further increasing the gap between (1.8) and (1.9).

However, for oversampling factor greater than two, *i.e.* $N/(2n) > 2$, the difference in using (1.8) or (1.9) becomes negligible. Both bounds coincide with the optimal value at $N = 4n$, and are within roughly 10% of the optimal value for large oversampling factors. Hence, both (1.8) and (1.9) are useful, in different oversampling regimes.

Next, we obtain a tighter estimate and a lower bound by restricting our attention to real polynomials.

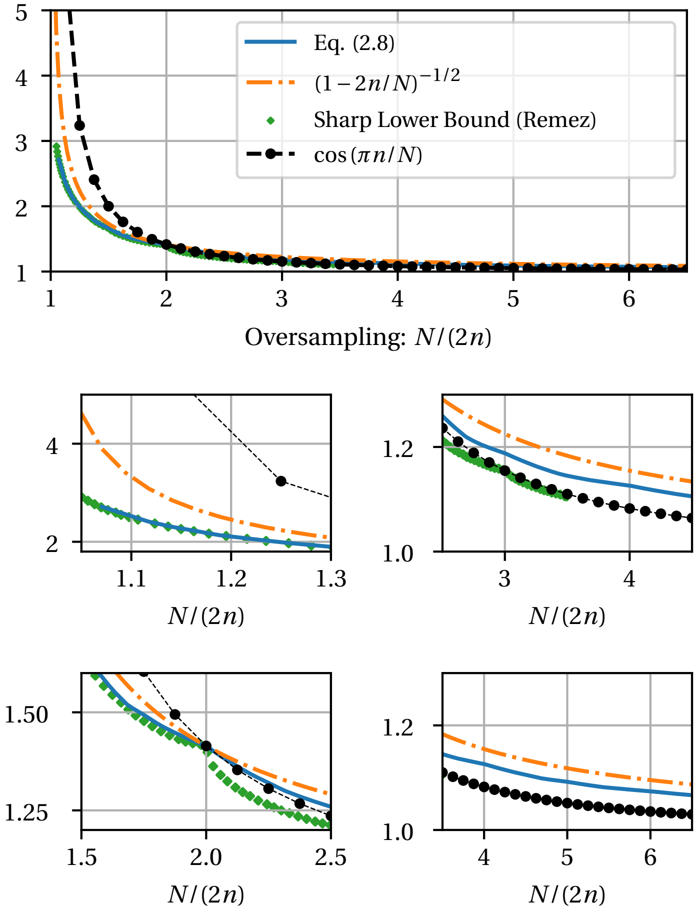


Figure 1.1: Comparing upper bounds of the form (1.7) as a function of oversampling ratio, $N/2n$, calculated with $n = 8$. Green diamonds indicate the optimal upper bound as calculated using a Remez-type algorithm [1, Fig. 2]. Black dots denote the upper bound (1.3) at valid locations, *i.e.* $N = 2m > 2n+1$.

Corollary 1.2. *Let $p \in \bar{T}_n^d$ and take $N \geq 2n+1$. Set $A \triangleq \max_{\omega\in\Theta_N^d} p(\omega)$, $B \triangleq \min_{\omega\in\Theta_N^d} p(\omega)$ and take $C_{N,n}^d$ as in Theorem 1.1. Then,*

$$p(\omega) \leq \frac{1}{2}\left(A + B + C_{N,n}^d\,(A-B)\right), \tag{1.10}$$

$$p(\omega) \geq \frac{1}{2}\left(A + B - C_{N,n}^d\,(A-B)\right), \tag{1.11}$$

$$\|p\|_\infty \leq \frac{1}{2}\left(|A+B| + C_{N,n}^d(A-B)\right). \tag{1.12}$$

The estimates (1.10) and (1.12) coincide with (1.7) in the case that

$$\min_{\omega\in\Theta_N^d} p(\omega) = -\max_{\omega\in\Theta_N^d} p(\omega),$$

and are tighter otherwise, making this refinement especially useful for non-negative polynomials.

By Theorem 1.1, $C_{N,n}^d \to 1$ as $N \to \infty$. Thus as $N \to \infty$, the right-hand side of (1.11) approaches B, and by continuity we have $B = \min_{\omega \in \Theta_N^d} p(\omega) \to \min_{\omega \in \mathbb{T}^d} p(\omega)$. Thus the bound is tight as $N \to \infty$. In the case of $A = B$, the right-hand side of (1.11) is $A = \|p\|_{N^d,\infty}$, and thus $p(\omega) > 0$ so long as the samples of p are not uniformly zero. This is expected, as otherwise the polynomial $p(\omega) - \|p\|_{N^d,\infty} \in T_n^d$ would vanish on a set of $N^d > (2n+1)^d$ points, which is impossible unless the polynomial is identically zero.

A little algebra on (1.11) establishes a sufficient condition to verify the strict positivity of a multivariate trigonometric polynomial.

Corollary 1.3. *Let $p \in \bar{T}_n^d$ and $N \geq 2n+1$. Set $\alpha = 2n/N$. If $p(\omega) > 0$ for all $\omega \in \Theta_N^d$ and*

$$\kappa_N \triangleq \frac{\max_{\omega \in \Theta_N^d} p(\omega)}{\min_{\omega \in \Theta_N^d} p(\omega)} < \frac{C_{N,n}^d + 1}{C_{N,n}^d - 1} \tag{1.13}$$

then $p(\omega) > 0$ for all $\omega \in \mathbb{T}^d$. Furthermore, as $C_{N,n}^d \leq (1-\alpha)^{-\frac{d}{2}}$, (1.13) can be replaced by the more stringent, but easier to evaluate, condition

$$\kappa_N < \frac{1 + (1-\alpha)^{\frac{d}{2}}}{1 - (1-\alpha)^{\frac{d}{2}}}. \tag{1.14}$$

For $p \in \bar{T}_n^d$ with non-negative samples, we call the quantity κ_N in (1.13) the *N-sample dynamic range.*

Corollary 1.3 provides an easy way to certify strict positivity of a real, non-negative polynomial from its samples: simply calculate the dynamic range κ_N and verify that (1.13) or (1.14) holds. These conditions are easier to satisfy (as a function of the oversampling rate) for polynomials whose maximum and minimum sampled values are close to one another. Intuitively, if the sampled values of a real trigonometric polynomial are strictly positive and don't vary "too much", then the polynomial is strictly positive over its entire domain. For fixed n and d, the right-hand side of (1.14) is an increasing function of N, illustrating a tradeoff: polynomials with a large amount of variation, and thus large values of κ_N, require larger oversampling factors N for the bounds to hold. Note that κ_N is not necessarily a monotone function of N, but is monotone in k when choosing $N = 2^k$. Figure 1.2 illustrates the regions for which (1.13) and (1.14) hold.

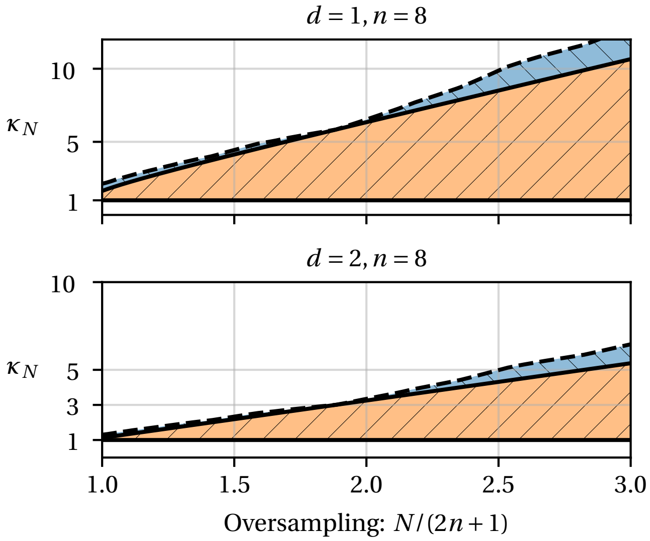


Figure 1.2: Any $p \in \bar{T}_n^d$ with positive samples and whose N-sample signed dynamic range κ_N lies in the shaded region must be strictly positive. The orange shaded region is certified using (1.14), while the blue region uses (1.13).

1.3 Proof of Theorem 1.1

We begin by proving Theorem 1.1, which extends the upper bound (1.5) from univariate to multivariate polynomials and provides a tighter result for the case of low oversampling. Due to the separable nature of T_n^d (*e.g.* T_n^d is the d-fold tensor product of T_n with itself), the proof is similar to the univariate case [1]. We consider both real and complex trigonometric polynomials.

1.3.1 Interpolation by the Dirichlet Kernel

For $\mathbf{n} = [n_1, \dots n_d] \in [N]^d$, the $\mathbf{n}$-th order Dirichlet kernel is the tensor product of d kernels, each of order n_i:

$$D_{\mathbf{n}}^d(\omega) \triangleq \sum_{|k_i| \le n_i} e^{jk\cdot\omega} = \prod_{i=1}^{d} \frac{\sin\frac{2n_i+1}{2}\omega_i}{\sin\frac{\omega_i}{2}} \quad \omega \in \mathbb{T}^d, k \in \mathbb{Z}^d. \tag{1.15}$$

If $\mathbf{n}$ is identical in each index (*i.e.* $n_i = n$ for each $i \in [d]$) we write the kernel as $D_n^d(\omega)$. The Dirichlet kernel is key to the periodic sampling formula.

Lemma 1.4. *Let $p \in T_n^d$ be sampled on Θ_N^d. Let m be an integer with $m \ge n$. If $N > n + m$, then*

$$p(\omega) = \frac{1}{N^d} \sum_{\omega_k \in \Theta_N^d} p(\omega_k) D_m^d(\omega - \omega_k) \tag{1.16}$$

for all $\omega \in \mathbb{T}^d$.

Lemma 1.4 (*e.g.*, [23]) is the periodic counterpart of sinc interpolation arising in the Whittaker-Shannon interpolation formula. The bound (1.2) can be obtained from (1.16) when $N = 2n+1$ [20].

1.3.2 Interpolation by the de la Vallée-Poussin Kernel

A better result is obtained by oversampling ($N > 2n+1$) and exploiting the nice properties of summation kernels.

Let n, m be integers with $m > n$ and define $\mathbb{V}_{n,m}^d = \left\{l \in \mathbb{Z}^d : n \le l_i < m\right\}$. The n,m-th de la Vallée-Poussin kernel is defined as the moving average of Dirichlet kernels:

$$\begin{aligned} D_{n,m}^d(\omega) &\triangleq \frac{1}{(m-n)^d} \sum_{\mathbf{n} \in \mathbb{V}_{n,m}^d} D_{\mathbf{n}}^d(\omega) \\ &= \frac{1}{(m-n)^d} \prod_{i=1}^d \frac{\sin\left(\frac{m+n}{2}\omega_i\right)\sin\left(\frac{m-n}{2}\omega_i\right)}{\sin^2(\omega_i/2)}. \end{aligned} \tag{1.17}$$

Taking $n = 0$ recovers the well-known Fejér kernel [24],

$$D_{0,m}^d = \frac{1}{m^d} \prod_{i=1}^d \frac{\sin^2\left(\frac{m}{2}\omega_i\right)}{\sin^2(\omega_i/2)}.$$

The Fejér kernel is used to derive the bound (1.4) [22].

Importantly, the de la Vallée-Poussin kernel inherits the reproducing property of the Dirichlet kernel.

Lemma 1.5. *For any* $p \in T_n^d$ *we have*

$$p(\omega) = \frac{1}{N^d} \sum_{\omega_k \in \Theta_N^d} p(\omega_k) D_{n,m}^d(\omega - \omega_k)$$

for all $\omega \in \mathbb{T}^d$ *whenever* $m > n$ *and* $N \ge n + m$.

Proof. Expanding the de la Vallée-Poussin kernel into a sum of Dirichlet kernels and applying

Lemma 1.4,

$$\begin{aligned}
&\frac{1}{N^d}\sum_{\omega_k\in\Theta_N^d} p(\omega_k)D_{n,m}^d(\omega-\omega_k)\\
&\quad=\frac{1}{(m-n)^d}\sum_{\mathbf{n}\in\mathbb{V}_{n,m}^d}\frac{1}{N^d}\sum_{\omega_k\in\Theta_N^d} p(\omega_k)D_{\mathbf{n}}^d(\omega-\omega_k)\\
&\quad=\frac{1}{(m-n)^d}\sum_{\mathbf{n}\in\mathbb{V}_{n,m}^d} p(\omega)=p(\omega).
\end{aligned}$$

□

1.3.3 Proof of Theorem 1.1

The upper bound of Theorem 1.1 depends on estimates of $\sum_{\omega_k\in\Theta_N^d}\left|D_{n,m}^d(\omega-\omega_k)\right|$, which we collect into a pair of lemmas.

Lemma 1.6. *Take $N\geq 2n+1$. Then, for all $\omega\in\mathbb{T}^d$,*

$$\begin{aligned}
&\sum_{\omega_k\in\Theta_N^d}\left|D_{n,N-n}^d(\omega-\omega_k)\right|\\
&\quad\leq\left(\sup_{\omega\in\mathbb{T}}\sum_{\omega_k\in\Theta_N}\left|D_{n,N-n}(\omega-\omega_k)\right|\right)^d \qquad (1.18)\\
&\quad=\frac{\left(\sup_{\omega\in\mathbb{T}}\left\{\sum_{\omega_k\in\Theta_N}\left|\frac{\sin\left(\frac{N\omega}{2}\right)\sin\left(\frac{N-2n}{2}(\omega-\omega_k)\right)}{\sin^2((\omega-\omega_k)/2)}\right|\right\}\right)^d}{(N-2n)^d}.
\end{aligned}$$

Proof. First, we fix notation: for $\omega_k\in\Theta_N^d$ and $k\in[N]^d$, we define $\omega_{k_i}=2\pi k_i/N$. Using (1.17), we have

$$\begin{aligned}
&\sum_{\omega_k\in\Theta_N^d}\left|D_{n,N-n}^d(\omega-\omega_k)\right|(N-2n)^d\\
&\quad=\sum_{\omega_k\in\Theta_N^d}\prod_{i=1}^d\left|\frac{\sin\left(\frac{N}{2}(\omega_i-\omega_{k_i})\right)\sin\left(\frac{N-2n}{2}(\omega_i-\omega_{k_i})\right)}{\sin^2((\omega_i-\omega_{k_i})/2)}\right|\\
&\quad\leq\left(\sup_{\omega\in\mathbb{T}}\sum_{\omega_k\in\Theta_N}\left|\frac{\sin\left(\frac{N}{2}(\omega-\omega_k)\right)\sin\left(\frac{N-2n}{2}(\omega-\omega_k)\right)}{\sin^2((\omega-\omega_k)/2)}\right|\right)^d \qquad (1.19)\\
&\quad=\left(\sup_{\omega\in\mathbb{T}}\sum_{\omega_k\in\Theta_N}\left|\frac{\sin\left(\frac{N\omega}{2}\right)\sin\left(\frac{N-2n}{2}(\omega-\omega_k)\right)}{\sin^2((\omega-\omega_k)/2)}\right|\right)^d,
\end{aligned}$$

where the final step follows from $\left|\sin\left(\frac{N}{2}(\omega - 2\pi k/N)\right)\right| = \left|\sin\left(\frac{N\omega}{2}\right)\right|$ for $k \in [N]$. The bound (1.18) is obtained by replacing (1.19) with the definition of $D_{n,N-n}(\omega)$ given by (1.17). □

The following lemma for univariate trigonometric polynomials is key to the derivation of (1.5). [3]

Lemma 1.7. *Let $m > n$ and take $N \geq n + m$. Then*

$$\sum_{\omega_k \in \Theta_N} \left|D_{n,m}(\omega - \omega_k)\right| \leq N \left(\frac{m+n}{m-n}\right)^{\frac{1}{2}}$$

for all $\omega \in \mathbb{T}$. In particular, taking $N \geq 2n+1$ and $m = N - n$ yields

$$\sum_{\omega_k \in \Theta_N} \left|D_{n,N-n}(\omega - \omega_k)\right| \leq N \left(\frac{N}{N-2n}\right)^{\frac{1}{2}}. \tag{1.20}$$

Proof. See [1, Theorem 1]. □

We are now set to complete the proof of Theorem 1.1.

Proof of Theorem 1.1. Without loss of generality, assume $\|p\|_{N^d,\infty} = 1$. Then, by Lemma 1.5, we have

$$\begin{aligned} \left|p(\omega)\right| &= \left|\frac{1}{N^d} \sum_{\omega_k \in \Theta_N^d} p(\omega_k) D^d_{n,N-n}(\omega - \omega_k)\right| \\ &\leq \frac{1}{N^d} \sum_{\omega_k \in \Theta_N^d} \left|p(\omega_k) D^d_{n,N-n}(\omega - \omega_k)\right| \qquad (1.21) \\ &\leq \frac{1}{N^d} \sum_{\omega_k \in \Theta_N^d} \left|D^d_{n,N-n}(\omega - \omega_k)\right|, \qquad (1.22) \end{aligned}$$

where (1.21) and (1.22) follow from the triangle inquality and Hölder's inequality, respectively.

Now, applying Lemma 1.6, we have

$$\begin{aligned} \left|p(\omega)\right| &\leq N^{-d} \left(\sup_{\omega \in \mathbb{T}} \sum_{\omega_k \in \Theta_N} \left|D_{n,N-n}(\omega - \omega_k)\right|\right)^d \qquad (1.23) \\ &= \frac{\left(\sup_{\omega \in \mathbb{T}} \left\{ \sum_{\omega_k \in \Theta_N} \left| \frac{\sin\left(\frac{N\omega}{2}\right) \sin\left(\frac{N-2n}{2}(\omega - \omega_k)\right)}{\sin^2\left((\omega - \omega_k)/2\right)} \right| \right\}\right)^d}{N^d (N-2n)^d}, \end{aligned}$$

[3] A multivariate extension is straightforward, but not used in the proof of Theorem 1.1 and is omitted here.

which implies (1.7)-(1.8). Applying the bound (1.20) of Lemma 1.7 to (1.23) yields

$$\left|p(\omega)\right| \leq \left(\frac{N}{N-2n}\right)^{\frac{d}{2}} = (1-\alpha)^{-\frac{d}{2}},$$

which establishes (1.9).

Finally, as $N \geq 2n+1$, by Taylor's theorem we have $(1-\alpha)^{-\frac{d}{2}} = 1 + \frac{dn}{N} + \mathcal{O}((dn/N)^2)$. It follows that

$$\begin{aligned} C_{N,n}^{d}\|p\|_{N^d,\infty} - \|p\|_\infty &\leq \left(C_{N,n}^{d} - 1)\right)\|p\|_\infty \\ &\leq \left((1-\alpha)^{-\frac{d}{2}} - 1\right)\|p\|_\infty \\ &= \left(\frac{dn}{N} + \mathcal{O}((dn/N)^2)\right)\|p\|_\infty, \end{aligned}$$

where we have used $\|p\|_{N^d,\infty} \leq \|p\|_\infty$. □

1.4 Proof of Refinement and Lower Bound for Real Trigonometric Polynomials

We now restrict our attention to real trigonometric polynomials. We will use the shorthand notation $A \triangleq \max_{\omega\in\Theta_N^d} p(\omega)$ and $B \triangleq \min_{\omega\in\Theta_N^d} p(\omega)$. Note both A and B are (not necessarily monotonic) functions of N.

The bound of Theorem 1.1 is at its tightest whenever the samples of $p(\omega)$ are centered about zero, *i.e.* $\min_{\omega\in\Theta_N^d} p(\omega) = -\max_{\omega\in\Theta_N^d} p(\omega)$, and can be loose otherwise. To see this, take $c > 0$ and consider the shifted polynomial $\tilde{p}(\omega) = p(\omega) + c$. Applying Theorem 1.1 yields

$$\begin{aligned} \|\tilde{p}\|_\infty &\leq C_{N,n}^{d}\|\tilde{p}\|_{N^d,\infty} \\ &\leq C_{N,n}^{d}(\|p\|_{N^d,\infty} + c). \end{aligned} \tag{1.24}$$

Applying the triangle inequality in advance of Theorem 1.1 results in

$$\|\tilde{p}\|_\infty \leq \|p\|_\infty + c \leq C_{N,n}^{d}\|p\|_{N^d,\infty} + c,$$

which may be much smaller than (1.24), but presupposes knowledge of c. While we do not know this offset, it can be estimated from the samples of $\tilde{p}$. This motivates our refined bound, Corollary 1.2, which we now prove.

Proof of Corollary 1.2. If $A = B$ then $p(\omega) - A$ vanishes on a set of $N^d \geq (2n+1)^d$ points; thus $p(\omega)$ is the constant polynomial $p(\omega) = A$ and (1.10) to (1.12) hold with equality.

Define $q \in T_n^d$ as $q(\omega) \triangleq p(\omega) - \frac{A+B}{2}$, which satisfies

$$\|q\|_{N^d,\infty} = \left| A - \frac{A+B}{2} \right| = \frac{A-B}{2}.$$

By Theorem 1.1, we have for all $\omega \in \mathbb{T}^d$,

$$\left| q(\omega) \right| \leq C_{N,n}^d \frac{A-B}{2}.$$

Combined with the definition of $q(\omega)$, we have

$$-C_{N,n}^d \frac{A-B}{2} \leq p(\omega) - \frac{A+B}{2} \leq C_{N,n}^d \frac{A-B}{2},$$

and rearranging gives (1.10) and (1.11).

Finally, we have

$$\begin{aligned} \left| p(\omega) \right| &\leq \left| q(\omega) \right| + \left| \frac{A+B}{2} \right| \\ &\leq C_{N,n}^d \frac{A-B}{2} + \left| \frac{A+B}{2} \right|, \end{aligned}$$

yielding (1.12). □

1.5 Examples

1.5.1 Univariate Example

Figure 1.3 illustrates our bounds for a randomly chosen univariate real trigonometric polynomial, $p \in \bar{T}_8^1$, given by[4]

$$\begin{aligned} p(\omega) \triangleq 3.9 + \frac{1}{2}\Big(&0.4\cos(\omega) + 1.0\sin(\omega) \\ &+ 2.2\cos(2\omega) + 1.9\sin(2\omega) - 1.0\cos(3\omega) + 1.0\sin(3\omega) \\ &- 0.2\cos(4\omega) - 0.1\sin(4\omega) + 0.4\cos(5\omega) + 0.1\sin(5\omega) \\ &+ 1.5\cos(6\omega) + 0.8\sin(6\omega) + 0.1\cos(7\omega) + 0.4\sin(7\omega) \\ &+ 0.3\cos(8\omega) + 1.5\sin(8\omega)\Big). \end{aligned} \tag{1.25}$$

Note that the bounds are not necessarily monotonic functions of N. We see that an oversampling factor of 1.3, or $N = 23$, is enough samples to certify the strict positivity of this polynomial.

1.5.2 Trivariate Example

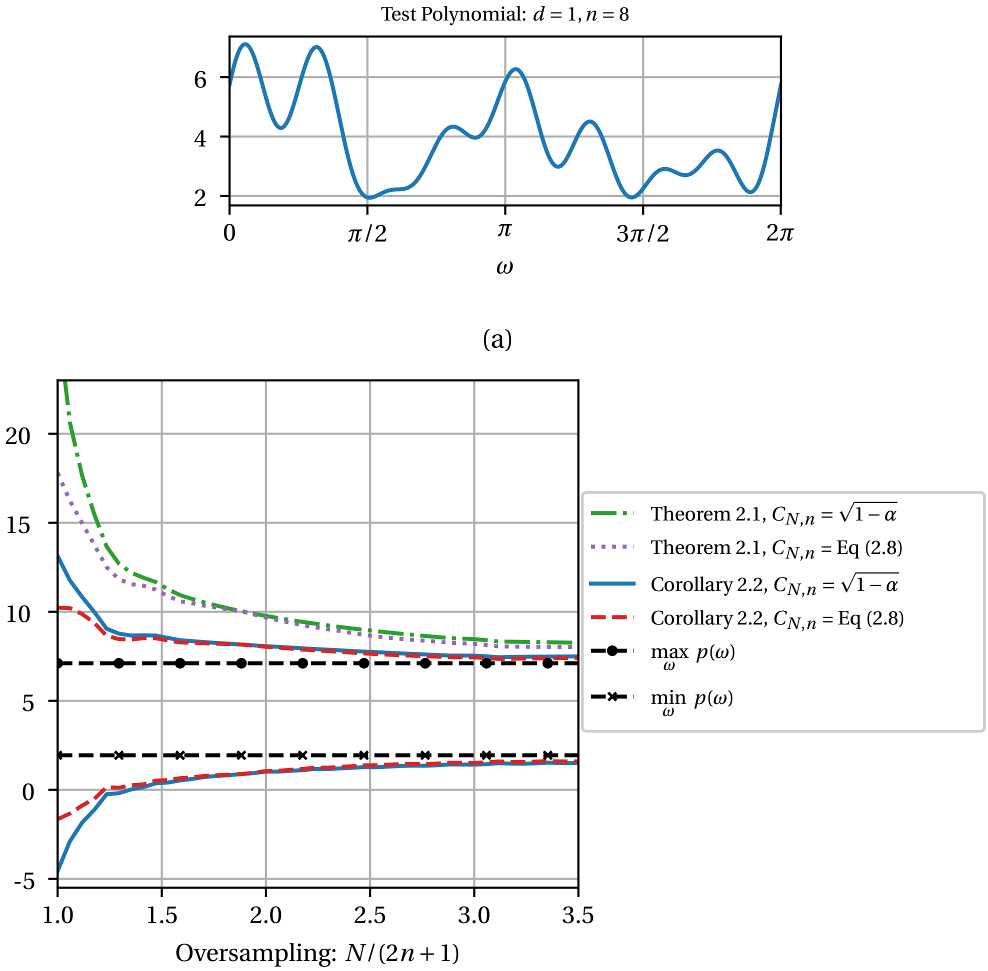
For simplicity, take $p \in \bar{T}_n^3$ to be $p(\omega) = D_n^3(\omega)/(2n+1)^3$, where D_n^3 is the Dirichlet kernel (1.15) with (uniform) degree n and the scaling is such that $\|p\|_\infty = 1$.

We obtain uniform samples of $p(\omega)$ over Θ_N^d by computing a zero-padded discrete Fourier transform. In particular, we embed an $n \times n \times n$ array of ones into an $N \times N \times N$ array of zeros, and apply the FFT algorithm to this array. We choose N to be a favorable size for the FFT algorithm, such as a power of two. As we choose N proportional to the degree n of p, our method scales as $\mathcal{O}(n^d \log n)$ with $d = 3$ in this example.

Figure 1.4 shows the estimates obtained using Corollary 1.2 as a function of N for a variety of orders n; the true maximum value of $p(\omega)$ is 1 and the minimum can be shown to be roughly $-2/(3\pi) \approx -0.22$. Evaluating the bounds for $n = 32$ and $N = 512$ took roughly one second on a workstation with an Intel i7-6700K CPU and 32GB of RAM.

To draw a comparison with the sum-of-squares framework, we use the `POS3POLY` MATLAB library, in particular the function `min_poly_value_multi_general_trig_3_5` [25]. This function finds the minimum value of a polynomial (given its coefficients) by a solving an SDP feasibility problem using an interior point method; the maximum value is obtained by calling the same function on $-p$. The per-iteration complexity of this method is $\mathcal{O}(n^{4d})$.

[4] The coefficients were drawn from a standard normal distribution and rounded to the first decimal point.



(a)

(b)

Figure 1.3: Example of upper and lower bounds for $p \in \bar{T}_8^1$ given by (1.25). (a): Test Polynomial. (b): Upper and lower bounds as a function of oversampling rate.

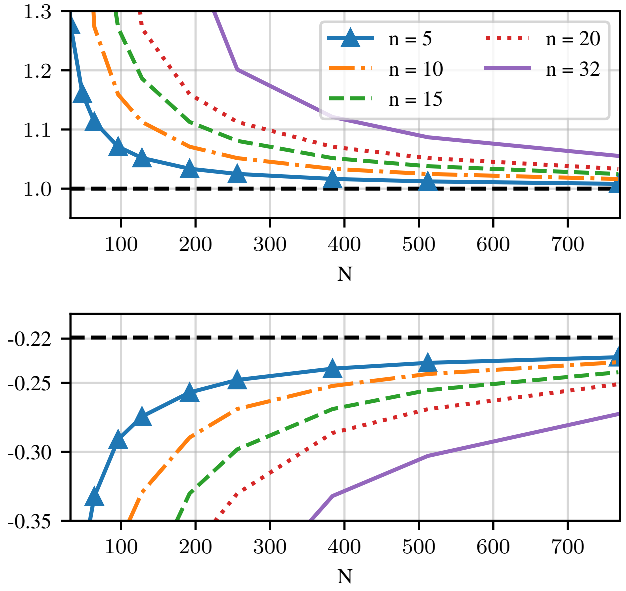


Figure 1.4: Upper and lower bounds for the Dirichlet kernel of 3 variables using Corollary 1.2.

For $n = 7$, `POS3POLY` required 75 seconds to obtain the minimum value to within 3×10^{-3}; $n = 8$ required 260 seconds and found the minimum to within 2×10^{-3}. The $n = 9$ case exhausted the system memory and was too large to solved on the workstation.

This is meant to be an illustrative, but certainly not exhaustive, comparison between the bounds presented in this chapter and the sum-of-squares framework. Sum-of-squares methods are especially attractive if an exact solution is needed or if the polynomial has sparse coefficients, in which case the complexity can be dramatically reduced.

1.6 Application to 2D Filter Bank Design

1.6.1 Perfect Reconstruction Filter Banks

We review a few key properties of multirate perfect reconstruction filter banks before turning to our design algorithm; see [8, 26] for a complete overview.

An N_c channel analysis filter bank operating on d-dimensional signals consists of a collection of N_c *analysis filters* h_i and a non-singular downsampling matrix $M \in \mathbb{Z}^{d \times d}$. A filter bank is *perfect reconstruction* (PR) if there exists a (possibly non-unique) synthesis filter bank, consisting of a collection of N_c *synthesis filters*, g_i, and the upsampling matrix M, that reconstructs a signal from its analyzed version. An analysis filter bank, along with its corresponding synthesis filter

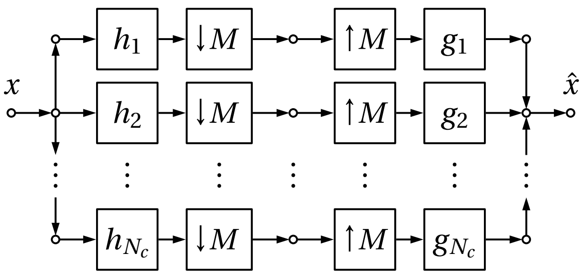


Figure 1.5: An N_c channel multi-rate filter bank with analysis filters h_i and synthesis filters g_i.

bank, are illustrated in Fig. 1.5. If the filter bank is PR then $\hat{x} = x$. In what follows, a "filter bank" indicates an analysis filter bank unless otherwise specified.

We consider finite impulse response (FIR) filters, and for simplicity, we restrict our attention to impulse responses with a square support. A real (square) d-variate (or d-dimensional) FIR filter h of length n is a function $h : \mathbb{Z}^d \to \mathbb{R}$ such that $h[m] = 0$ if $m_i < 0$ or $m_i \geq n$ for any $0 \leq i < d$.

A multidimensional discrete-time signal is a function $x : \mathbb{Z}^d \to \mathbb{R}$. Downsampling a signal x by a non-singular integer matrix M retains only the samples on the lattice generated by M; that is, integer vectors of the form $v = Mt$. The simplest choice of downsampling matrix is $M = sI_d$, where the integer $s \geq 1$ controls the downsampling factor and I_d is the identity matrix in d dimensions. We will refer to this as the *uniform* downsampling scheme.

The i-th polyphase component of a signal x is a function $\hat{x}^i : \mathbb{Z}^d \to \mathbb{R}$ obtained by shifting and downsampling x. In particular, $\hat{x}^i[m] = x[Mm + v_i]$ for $m \in \mathbb{Z}^d$, where v_i is an integer vector of the form Mt and $t \in [0,1)^d$. There are $|M| \triangleq \det M$ such integer vectors, and each generates one polyphase component of the signal. The z-transform of the i-th polyphase component of x is $\hat{X}^i(z) = \sum_{n \in \mathbb{Z}^d} x[Mn + v_i] z^{-n}$, where $z \in \mathbb{C}^d$ and $z^{-n} = z_1^{-n_1} z_2^{-n_2} \ldots z_d^{-n_d}$.

The polyphase decomposition of an analysis filter is defined in a similar fashion. The i-th polyphase component of the analysis filter h is $\hat{h}^k[m] = h[Mm - v_i]$; note the difference in sign when compared to the definition of $\hat{x}^i$.

A d-dimensional filter bank with filters $\{h_i\}_{i=1}^{N_c}$ and downsampling matrix M has a polyphase matrix $\hat{\mathbf{H}}(z) \in \mathbb{C}^{N_c \times |M|}$ formed by stacking the polyphase components of each analysis filter into a row vector, and stacking the N_c rows into a matrix. Explicitly,

$$\hat{\mathbf{H}}(z) \triangleq \begin{bmatrix} \hat{H}_0^0(z) & \hat{H}_0^1(z) & \ldots & \hat{H}_0^{|M|-1}(z) \\ \hat{H}_1^0(z) & \hat{H}_1^1(z) & \ldots & \hat{H}_1^{|M|-1}(z) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{H}_{N^c-1}^0(z) & \hat{H}_{N^c-1}^1(z) & \ldots & \hat{H}_{N_c-1}^{|M|-1}(z) \end{bmatrix},$$

where $\hat{H}_i^k(z)$ is the z-transform of the k-th polyphase component of the i-th filter. The entries of

$\hat{\mathbf{H}}(z)$ are multi-variate Laurent polynomials in $z \in \mathbb{C}^d$ and become trigonometric polynomials when restricted to the unit circle; that is, $z = e^{j\omega}$ with $\omega \in \mathbb{T}^d$. In a customary abuse of notation, we write $\hat{\mathbf{H}}(\omega) \triangleq \hat{\mathbf{H}}(e^{j\omega})$.

There are deep connections between perfect reconstruction filter banks and redundant signal expansions using *frames* [12, 27–29]. In particular, oversampled perfect reconstruction filter banks implement an *frame expansion*. Associated with a perfect reconstruction filter bank are a pair of scalars, the *upper and lower frame bounds*, defined by

$$A \triangleq \text{ess sup}_{\omega \in \mathbb{T}^d, m=1,\ldots |M|} \lambda_n(\omega),$$
$$B \triangleq \text{ess inf}_{\omega \in \mathbb{T}^d, m=1,\ldots |M|} \lambda_n(\omega),$$

where $\lambda_n(\omega)$ is an eigenvalue of the matrix $\hat{\mathbf{H}}^*(\omega)\hat{\mathbf{H}}(\omega)$. If $A = B$ the frame is said to be *tight*. The ratio A/B is the *frame condition number*; if $A/B \approx 1$, the frame is said to be *well-conditioned*. The frame bounds of a filter bank determine important numerical properties such as sensitivity to perturbations, and the frame condition number serves a similar role as the condition number of a matrix.

The synthesis filter bank also admits a polyphase decomposition. The i-th polyphase component of a synthesis filter g is $\hat{g}^k[m] = g[Mm + v_i]$. The synthesis polyphase matrix is of size $|M| \times N_c$ and has entries

$$\hat{\mathbf{G}}(z) \triangleq \begin{bmatrix} \hat{G}_0^0(z) & \hat{G}_1^0(z) & \cdots & \hat{G}_{|M|-1}^0(z) \\ \hat{G}_0^1(z) & \hat{G}_1^1(z) & \cdots & \hat{G}_{|M|-1}^1(z) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{G}_0^{N^c-1}(z) & \hat{G}_1^{N^c-1}(z) & \cdots & \hat{G}_{|M|-1}^{N_c-1}(z) \end{bmatrix}.$$

If a pair of analysis and synthesis filter banks share the PR property, then $\hat{\mathbf{G}}(z)\hat{\mathbf{H}}(z) = I_{|M|}$, where $I_{|M|}$ is the $|M| \times |M|$ identity matrix. That is, $\hat{\mathbf{G}}(z)$ is a left inverse for $\hat{\mathbf{H}}(z)$. If $N_c > |M|$, the filter bank is said to be *oversampled*, and the synthesis filter bank is not unique. A particular choice is the *minimum-norm synthesis filter bank*, given by

$$\hat{\mathbf{H}}^{\dagger}(z) \triangleq \left(\tilde{\mathbf{H}}(z)\hat{\mathbf{H}}(z)\right)^{-1}\tilde{\mathbf{H}}(z), \tag{1.26}$$

where the *para-conjugate* matrix $\tilde{\mathbf{H}}(z)$ is obtained by conjugating the polynomial coefficients of $\hat{\mathbf{H}}(z)$, replacing the argument z by z^{-1}, and transposing the matrix. On the unit circle, $\hat{\mathbf{H}}^{\dagger}(\omega) = \left(\hat{\mathbf{H}}^*(\omega)\hat{\mathbf{H}}(\omega)\right)^{-1}\hat{\mathbf{H}}^*(\omega)$.

A filter bank is perfect reconstruction if and only if its polyphase matrix has full column rank on the unit circle [6,12]. As the matrix $\hat{\mathbf{H}}^*(\omega)\hat{\mathbf{H}}(\omega)$ is positive semidefinite, the perfect reconstruction property holds if and only if the trigonometric polynomial

$$p_H(\omega) \triangleq \det\left(\hat{\mathbf{H}}^*(\omega)\hat{\mathbf{H}}(\omega)\right)$$

is strictly positive. This property is key to our proposed filter bank design algorithm.

The degree of $p_H(\omega)$ depends on the filter length and the downsampling matrix. To illustrate, we bound from above the degree of $p_H(\omega)$ when using separable downsampling. After downsampling by $M = sI_d$, a FIR filter of length n retains at most ceil(n/s) entries along each dimension; thus the polyphase component $\hat{H}_i^k(\omega)$ has maximum component degree $n' \triangleq \text{ceil}(n/s) - 1$. Note that $\hat{H}_i^k(\omega)$ contains only negative powers of ω; that is,

$$\hat{H}_i^k(\omega) \in \operatorname{span}\left\{e^{-jk\cdot\omega} : \omega \in \mathbb{T}^d, k \in \mathbb{Z}^d,\ n' \le k_i \le 0\right\}.$$

As such, the trigonometric polynomials $(\hat{H}_i^k(\omega))^* \hat{H}_i^l(\omega)$ remain in $T_{n'}^d$ and the entries of the matrix $\hat{\mathbf{H}}^*(\omega)\hat{\mathbf{H}}(\omega)$ are in the same space.

At worst, the determinant includes the product of $|M| = s^d$ polynomials of degree n', and so $p_H \in \bar{T}_m^d$ with

$$m \le s^d(\text{ceil}(n/s) - 1). \tag{1.27}$$

Taking $n = 12, s = 2$ and $d = 2$, we have $p_H \in T_{20}^2$.

1.6.2 Filter Bank Design: Analysis

The simplest multidimensional PR filter banks apply a 1D PR filter bank independently to each signal dimension; for example, in 2D, to the horizontal and vertical directions. These *separable* filters are written as a product of multiple 1D filters and suffer from limited directional sensitivity. The design and construction of *non-separable* multidimensional filter banks is difficult due to the lack of a spectral factorization theorem [9]; indeed, directly verifying the perfect reconstruction condition for a 2D filter bank is equivalent to determining the minimum value of a trigonometric polynomial and is thus NP-hard [3,4].

Some 2D PR filter banks, such as curvelets, have been hand-designed [30,31]. Other design methods include variable transformations applied to a 1D PR filter bank [8,26], modulating a prototype filter [26], invoking tools from algebraic geometry [10], or by solving an optimization problem [32,33].

Optimizing a filter bank subject to the PR condition is a semi-infinite optimization problem: we have a finite number of design variables, namely the filter coefficients, and the resulting polyphase matrix must be positive semidefinite over $\mathbb{T}^d$.

One approach is to carefully parameterize the filter bank architecture in such a way that guarantees the PR property [9,33]. A different approach is to relax the PR condition to *near* PR, and minimize the resulting reconstruction error using an iterative algorithm [32].

We use a different approach: we relax the semi-infinite problem into a finite one, then use Corollary 1.3 to certify that the solution of the relaxed problem is also a solution to the original problem. In particular, we design the filter bank such that $p_H(\omega)$ is strictly positive over the finite collection of sampling points Θ_N^d. Corollary 1.3 tells us that if the bounds (1.13) or (1.14) are satisfied, then $p_H(\omega)$ is strictly positive over all of $\mathbb{T}^d$, and the filter bank is thus PR.

We design our filter banks with an eye toward the bounds of Corollary 1.3: we want the maximum and minimum sampled values of $p_H(\omega)$ to be close to one another, so that the bounds (1.13) and (1.14) are satisfied for smaller values of N.

Our filter design approach is highly flexible. It applies to arbitrary filter lengths, any nonsingular decimation matrix, and will design PR filter banks in any number of dimensions. For simplicity we focus on designing real, 2D filter banks ($d = 2$) but our approach can be modified for d-dimensional complex filters.

We begin by specifying the number of channels, N_c, downsampling matrix M, and filter size. We require that $N_c \geq |M|$ so that the PR condition can hold. For simplicity, we use downsampling of the form $M = sI_2$, but our method can design filter banks using non-separable (*e.g.*, quincunx) downsampling matrices. We also constrain each filter to be of size $n \times n$, although this can be easily relaxed.

With these parameters set, we calculate the maximum degree m of $p_H(\omega)$ using (1.27). Next, we select the number of sampling points, N, to use during the design process. The conditions of Corollary 1.3 require we take $N \geq 2m+1$, but in practice we take $N > 4m$ so that we can tolerate larger values of κ_N while still certifying the perfect reconstruction property.

The i-th $n \times n$ filter will be written h_i, and we group the filters into a tensor $H \in \mathbb{R}^{N_c \times n \times n}$. The Discrete-Time Fourier Transform of the i-th filter is

$$h_i(\omega) = \sum_{m \in [n]^2} h_i[m] e^{j\omega \cdot m} \quad \omega \in \mathbb{T}^2,$$

and the squared magnitude response of h_i is $|h_i(\omega)|^2$.

Our goal is to design a perfect reconstruction filter bank where the magnitude response of the i-th channel matches a desired real and non-negative magnitude response $D_i(\omega)$ for $\omega \in \mathbb{T}^2$. We use a weighted quadratic penalty that measures the discrepancy between the magnitude response of a candidate filter and the D_i at the 2D-DFT samples Θ_N^2. Our filter design function is written

$$f(H,D) \triangleq \sum_{i=1}^{N_c} \sum_{\omega \in \Theta_N^2} W_i(\omega) \cdot \left| \left| \hat{h}_i(\omega) \right|^2 - D_i(\omega) \right|^2,$$

where we have introduced weighting functions $W_i(\omega)$ to control the importance given to the passband, transition band, and stop band. If D_i is not specified for some i, we take $W_i(\omega)$ to be uniformly zero; then h_i does not contribute to $f(H,D)$ but may contribute to the PR property of the filter bank.

We emphasize that other choices of a design function are possible; for instance, one could use a minimax criterion and minimize the maximum deviation between $\hat{h}_i(\omega)$ and $D_i(\omega)$. In Chapter 2 of this thesis we use a similar approach to learn signal-adapted undecimated perfect reconstruction (analysis) filter banks under a sparsity-inducing criterion [34].

In some cases, the filter design function alone may promote perfect reconstruction filter banks—for instance, when designing a non-decimated ($M = I_d$) filter bank where the desired magnitude responses satisfy a partition-of-unity condition. In general, though, this term is not enough. We add an additional regularization term to encourage filter banks that can be certified as perfect reconstruction using Corollary 1.3. Our regularizer is given by

$$R(H) \triangleq \alpha \sum_{i=1}^{N_c} \|h_i\|_F^2 + \sum_{\omega \in \Theta_N^2} \beta p_H(\omega)^2 - \gamma \log p_H(\omega),$$

where the non-negative scalars α, β, γ are tuning parameters. The first term prohibits the filter norms from becoming too large. The second and third terms apply the function $\omega \mapsto p_H(\omega)^2 - \log p_H(\omega)$ for each $\omega \in \Theta_N^2$. The negative logarithm barrier function becomes large when $p_H(\omega)$ goes to zero and the quadratic part discourages large values of $p_H(\omega)$.

Together, these terms ensure the matrix $\hat{\mathbf{H}}(\omega)$ is left invertible and well-conditioned for each $\omega \in \Theta_N^2$. They also ensure $p_H(\omega)$ does not grow too large over the sampling set. These properties ensure $p_H(\omega)$ is strictly positive and does not vary too much over Θ_N^2; thus, by Corollary 1.3, $R(H)$ promotes well-conditioned perfect reconstruction filter banks. We emphasize that this regularizer, as well as the filter design function, are only computed over on the discrete set Θ_N^2; passage to the continuous case is handled by Corollary 1.3.

Our designed filter bank is the solution to the optimization problem

$$\min_{H \in \mathcal{C}} f(H,D) + R(H),$$

where the constraint set $\mathcal{C}$ reflects any additional constraints on the filters, *e.g.* symmetry.

This minimization can be solved using standard first order methods such as gradient descent.

The main challenge is calculating the gradient of $\log(p_H(\omega))$, which is unwieldy for all but the shortest filters. A finite-difference approximation to the gradient can suffice, but we have had success using the reverse-mode automatic differentiation capabilities of the `autograd`[5] and `Pytorch`[6] Python packages. Our algorithm is implemented in `Pytorch` and runs on an NVidia Titan X GPU.

1.6.3 Experiment: Design of a Curvelet-Like Filter Bank

Our goal is to design a filter bank that approximates the discrete curvelet filter bank. Our desired magnitude responses are obtained from the frequency space tiling illustrated in Fig. 1.6; each channel should have a pass-band corresponding to a cell in this tiling. As the magnitude frequency response of a real filter is symmetric, *e.g.* $\left|\hat{h}(\omega_1,\omega_2)\right| = \left|\hat{h}(-\omega_1,-\omega_2)\right|$, 17 filters are needed for the desired partitioning. We use uniform downsampling by a factor of 2, that is, $M = 2I_2$. The filter bank is roughly 4× oversampled.

The weighting functions $W_i(\omega)$ were set to 1. We set $\beta = 10$ and $\alpha = \gamma = 1$. We used 5000 iterations of the Adam optimization algorithm with a learning rate of 10^{-2} [35]. The optimization completed in under one minute for all tasks.

We designed two filter banks; one with 8×8 filters and the other with 11×11 filters. We used $N = 64$ for both cases. The final filter banks and their magnitude responses are shown in Fig. 1.7.

We tested two methods to initialize the algorithm. In the first method, we take an $N \times N$ inverse DFT of the desired magnitude response, D_i, and extract the $n \times n$ central region of the resulting impulse response. Our second method is a simple random initialization. Both methods perform equally well in our design task.

We use Corollary 1.3 to verify the final filter banks are perfect reconstruction. For our filter bank with 8×8 filters, the bound (1.27) indicates $p_H \in \bar{T}^2_{12}$. Our sufficient condition in Corollary 1.3 for strict positivity requires $\kappa_{64} \le 4.4$, with κ_N given by (1.13). We computed $p_H(\omega)$ over all points in Θ^2_{64}, and used these values to compute κ_{64}. We found $\kappa_{64} = 1.3$ for the designed filter bank, and thus the filter bank is perfect reconstruction. When using 11×11 filters, we have $p_H \in \bar{T}^2_{20}$. This filter bank too is perfect reconstruction, as $\kappa_{64} = 1.8 \le 2.2$.

[5] `https://github.com/HIPS/autograd`

[6] `http://pytorch.org/`

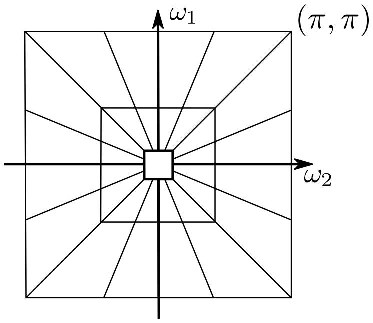


Figure 1.6: Desired tiling of frequency space.

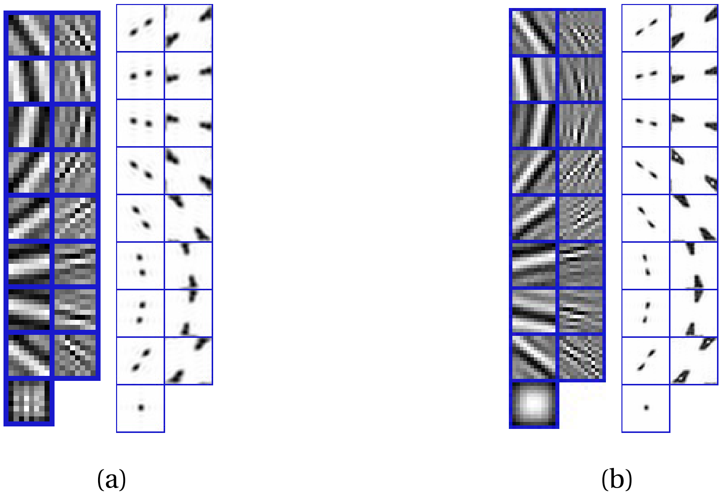


Figure 1.7: Optimized 17 channel filter bank. The left column of each subfigure shows the filter impulse response. The right column shows the magnitude frequency response, with $\omega = 0$ located at the center of each blue box. (a) 17 channel filter bank with 8×8 filters. (b) 17 channel filter bank with 11×11 filters.

1.6.4 Filter Bank Design: Synthesis

Our filter design problem has focused exclusively on the analysis portion of the filter bank, but in many applications the synthesis filter bank is equally important.

We focus on the oversampled case, *i.e.* $N_c > |M|$. The choice of synthesis filter bank is not unique. We have already seen one possible choice- the minimum-norm synthesis filter bank (1.26), which can be obtained explicitly once the analysis filter bank has been designed. In general, the minimum-norm synthesis filter bank consists of infinite impulse response (IIR) filters [12, 36].

In many applications, IIR filters are not practical—only FIR filters can be used, and short FIR filters are especially desirable from a computational perspective.

Fortunately, the redundancy of an oversampled filter bank affords us design flexibility. Sharif investigated when a generic[7] one-dimensional oversampled PR analysis filter bank admits a synthesis filter bank with short FIR filters. He found that almost all sufficiently oversampled PR analysis filter banks have such a synthesis filter bank, and obtained bounds on the minimum synthesis filter length [37]. The bounds depend only on the number of channels, downsampling factor, and analysis filter length, but not on the filter coefficients themselves.

We have a few options if a FIR synthesis filter bank is desired. The simplest solution is to truncate the (IIR) minimum-norm synthesis filters to a particular length. Indeed, a well-conditioned PR analysis filter bank has minimum-norm synthesis filters with coefficients that exhibit decay exponentially with filter length, implying that the minimum-norm synthesis filter bank can be well-approximated by FIR filters [38].

A second option is to use tools from algebraic geometry to find an FIR synthesis filter bank, if one exists [39].

We adopt a third option: we incorporate the desire for an FIR synthesis filter bank directly into the design problem. We add an additional set of FIR filters, denoted $\{g_i\}_{i=1}^{N_c}$, to the design parameters. The synthesis filters need not be the same length as the analysis filters. Our goal is for the polyphase matrix associated with the synthesis filter bank, $\hat{\mathbf{G}}(\omega)$, to be a left inverse of the analysis polyphase matrix. This condition is represented by the constraint

$$\hat{\mathbf{G}}(\omega)\hat{\mathbf{H}}(\omega) = I_{|M|}. \tag{1.28}$$

In practice, we solve an unconstrained problem using the quadratic penalty method: we penalize the distance between $\hat{\mathbf{G}}(\omega)\hat{\mathbf{H}}(\omega)$ and $I_{|M|}$ for each $\omega \in \Theta_N^2$ using the Frobenius norm [40]. Our

[7] A "generic" filter bank is one that is drawn at random; *i.e.* not a pathological choice.

modified design problem is given by

$$\min_{H,G\in\mathcal{C}} f(H,D) + R(H) + \lambda \sum_{\omega\in\Theta_N^2} \|\hat{\mathbf{G}}(\omega)\hat{\mathbf{H}}(\omega) - I_{|M|}\|_F^2. \tag{1.29}$$

We again use a first-order method, but now increase λ as a function of the iteration number so as to ensure $\hat{\mathbf{G}}(\omega)$ is a left inverse of $\hat{\mathbf{H}}(\omega)$.

As before, our new regularizer is evaluated only over Θ_N^2, not $\mathbb{T}^2$. For fixed, finite filter lengths, the entries of $\hat{\mathbf{G}}(\omega)\hat{\mathbf{H}}(\omega)$ are real trigonometric polynomials of bounded degree, and we can use the bounds of Corollary 1.2 to either ensure the constraint (1.28) holds over $\mathbb{T}^2$ or to estimate and bound the amount that the constraint has been violated.

1.6.5 Experiment: Filter Bank Design with FIR Synthesis Filters

We repeat the design experiment from Section 1.6.3 using the new objective function (1.29). As before, we use 17 channels and take $M = 2I_2$, leading to a roughly 4× oversampled filter bank. We work with 11×11 filters. We used 5000 iterations of the Adam optimization algorithm with a learning rate of 10^{-2}, and set the parameter $\lambda := \log_2(i)$ at iteration i.

Figure 1.8 collects the design results. Figure 1.8a shows the 11×11 analysis filters embedded into a larger 40×40 region. This is done to facilitate comparison with the minimum-norm synthesis filters, shown in Fig. 1.8b. The minimum-norm synthesis filters exhibit fast decay, as expected for a well-conditioned filter bank. The designed FIR synthesis filters, the $\{g_i\}_{i=1}^{N_c}$, are shown in Fig. 1.8c. These filters have no discernible structure. However, we computed $\|\hat{\mathbf{G}}(\omega)\hat{\mathbf{H}}(\omega) - I_{|M|}\|_F^2 < 10^{-7}$ for each $\omega \in \Theta_{128}^2$, this is a synthesis filter bank for $\hat{\mathbf{H}}$. Indeed, passing the standard `barbara` test image through the pair of analysis and synthesis filter banks yielded a reconstruction peak signal to noise ratio (PSNR) of more than 80 dB.

Figure 1.9 illustrates the coefficient decay properties of the minimum-norm synthesis filters. We show the square root of the absolute value of the filter coefficients to compress the dynamic range of the image. We see the expected exponential decay of filter coefficients associated with a well-conditioned filter bank [38].

1.7 Conclusion

We have proposed a fast and simple method to estimate the extremal values of a multivariate trigonometric polynomial directly from its samples. We have extended an existing upper bound from univariate to multivariate polynomials, and developed a strengthened upper bound and

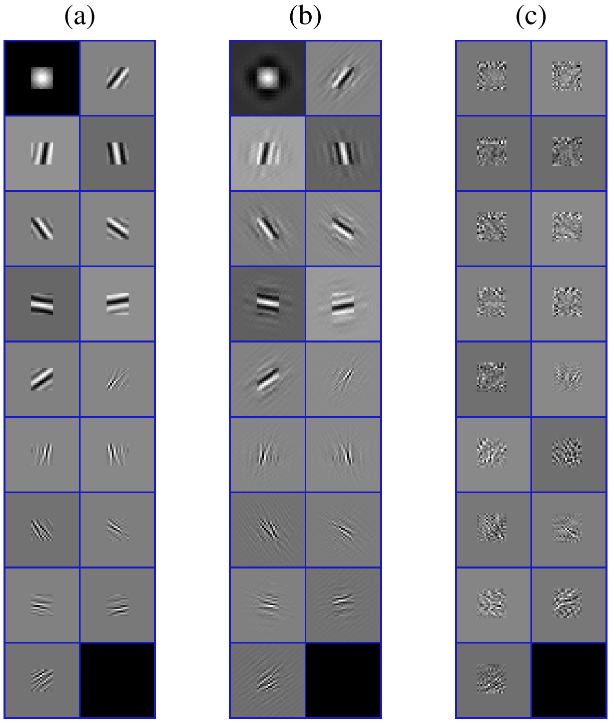


Figure 1.8: Analysis and synthesis filters for filter bank designed in Section 1.6.5. (a) Designed 11 × 11 analysis filters embedded into 40 × 40 filter. (b) Minimum-norm synthesis filters, obtained using (1.26). The filters exhibit fast coefficient decay; see Fig. 1.9. (c) Designed 16 × 16 FIR synthesis filters.

new lower bound for real trigonometric polynomials. The lower bound provides a new sufficient condition to certify global positivity of a real multivariate trigonometric polynomial.

We applied these results to the design of multidimensional perfect reconstruction filter banks, and demonstrated the construction of filter banks with both FIR and IIR synthesis filters. Future work will apply these results to the design of data-adaptive sparsifying filterbanks.

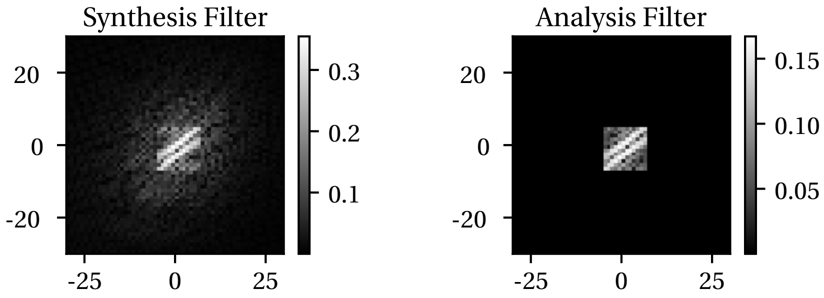


Figure 1.9: Square-root of absolute value of filter coefficients from one of the filters in Fig. 1.8a. Left: Minimum-norm synthesis filter exhibits fast coefficient decay, can be approximated with FIR filter. Right: FIR analysis filter.

Chapter 2

Learning Filter Bank Sparsifying Transforms

2.1 Introduction

Countless problems, from statistical inference to geological exploration, can be stated as the recovery of high-quality data from incomplete and/or corrupted linear measurements. Often, recovery is possible only if a model of the desired signal is used to regularize the recovery problem.

A powerful example of such a signal model is the *sparse representation*, wherein the signal of interest admits a representation with few non-zero coefficients. Sparse representations have traditionally been hand-designed for optimal properties on a mathematical signal class, such as the coefficient decay properties of a cartoon-like signal under a curvelet representation [30]. Unfortunately, these signal classes do not include the complicated and textured signals common in applications; further, it is difficult to design optimal representations for high-dimensional data. In light of these challenges, methods to *learn* a sparse representation, either from representative training data or directly from corrupted data, have become attractive.

We focus on a particular type of sparse representation, called *transform sparsity*, in which the signal $x \in \mathbb{R}^N$ satisfies $Wx = z + \eta$. The matrix $W \in \mathbb{R}^{K \times N}$ is called a *sparsifying transform* and earns its name as $z \in \mathbb{R}^K$ is sparse and $\|\eta\|_2$ is small [41]. Of course, a W that is uniformly zero satisfies this definition but provides no insight into the transformed signal. Several algorithms have been proposed to learn a sparsifying transform from data, and each must contend with this type of degenerate solution. The most common approach is to ensure that W is left invertible, so that Wx is uniformly zero if and only if x is uniformly zero. Such a matrix is a *frame* for $\mathbb{R}^n$.

In principle, we can learn a sparse representation for any data represented as a vector, including data from genomic experiments or text documents, yet most research has focused on learning models for spatio-temporal data such as images. With these signals it is common to learn a model for smaller, possibly overlapping, blocks of the data called *patches*. We refer to this type of model as a *patch-based* model, while we call a model learned directly at the image level an *image-based* model. Patch-based models tend to have fewer parameters than an unstructured image-based model, leading to lower computational cost and reduced risk of overfitting. In addition, an image contains many overlapping patches, and thus a model can be learned from a single noisy image [42].

Patch-based models are not without drawbacks. Any patch-based W learned using the usual frame constraints must have at least as many rows as there are elements in a patch, *i.e.* W must be square or tall. This limits practical patch sizes as W must be small to benefit from a patch-based model.

If our ultimate goal is image reconstruction, we must be mindful of the connection between extracted patches and the original image. Requiring W to be a frame for patches ignores this relationship and instead requires that each patch can be independently recovered. Yet, neighboring patches can be highly correlated—leading us to wonder if the patch-based frame condition is too strict. This leads to the question at the heart of this chapter: *Can we learn a sparsifying transform that forms a frame for images, but not for patches, while retaining the computational efficiency of a patch-based model?*

In this chapter, we show that existing sparsifying transform learning algorithms can be viewed as learning perfect reconstruction filter banks. This perspective leads to a new approach to learn a sparsifying transform that forms a frame over the space of images, and is structured as an undecimated, multidimensional filter bank. We call this structure a *filter bank sparsifying transform*. We keep the efficiency of a patch-based model by parameterizing the filter bank in terms of a small matrix W. In contrast to existing transform learning algorithms, our approach can learn a transform corresponding to a tall, square, or fat W. Our learned model outperforms earlier transform learning algorithms while maintaining low cost of learning the filter bank and the processing of data by it. Although we restrict our attention to 2D images, our technique is applicable to any data amenable to patch-based methods, such as 3D imaging data.

The rest of the chapter is organized as follows. In Section 2.2 we review previous work on transform learning, analysis learning, and the relationship between patch-based and image-based models. In Section 2.3 we develop the connection between perfect reconstruction filter banks and patch-based transform learning algorithms. We propose our filter bank learning algorithm in Section 2.4, describe denoising algorithms in Section 2.5, and present numerical results in Section 2.6. In Section 2.7 we compare our learning framework to the current crop of deep learning inspired approaches, and conclude in Section 2.8.

2.2 Preliminaries

2.2.1 Notation

Matrices are written as capital letters, while general linear operators are denoted by script capital letters such as $\mathcal{A}$. Column vectors are written as lower-case letters. The i-th component of a vector x is x_i. The i,j-th element of a matrix A is A_{ij}. We write the j-th column of A as $A_{:,j}$, and $A_{i,:}$ is the column vector corresponding to the transpose of the i-th row. The transpose and Hermitian transpose are A^T and A^*, respectively. Similarly, $\mathcal{A}^*$ is the adjoint of the linear operator $\mathcal{A}$. The $n \times n$ identity matrix is I_n. The n-dimensional vectors of ones and zeros are written as $\mathbf{1}_n$ and $\mathbf{0}_n$, respectively. For $x \in \mathbb{R}^N$, the diagonal matrix $\operatorname{ddiag}(x) \in \mathbb{R}^{N\times N}$ has the entries of x along its diagonal. The convolution of signals x and y is written $x * y$. For vectors $x, y \in \mathbb{R}^N$, the Euclidean inner product is $\langle x, y\rangle = \sum_{i=1}^N x_i y_i$ and the Euclidean norm is written $\|x\|_2$. The vector space of matrices $\mathbb{R}^{M\times N}$ is equipped with the inner product $\langle X, Y\rangle = \operatorname{trace}\left(X^T Y\right)$; the Frobenius norm is written $\|X\|_F$. When necessary, we indicate the vector space on which the inner product is defined; *e.g.* $\langle x, y\rangle_{\mathbb{R}^N}$.

2.2.2 Transform Sparsity

Recall that a signal $x \in \mathbb{R}^N$ satisfies the transform sparsity model if there is a matrix $W \in \mathbb{R}^{K\times N}$ such that $Wx = z + \eta$, where z is sparse and $\|\eta\|_2$ is small. The matrix W is called a *sparsifying transform* and the vector z is a *transform sparse code.* Given a signal x and sparsifying transform W, the *transform sparse coding* problem is

$$\operatorname*{arg\,min}_z \frac{1}{2}\|Wx - z\|_2^2 + \nu\psi(z) \tag{2.1}$$

for a sparsity promoting functional ψ. Exact s-sparsity can be enforced by selecting ψ to be the indicator function over the set of s-sparse vectors. We recognize (2.1) as the evaluation of the proximal operator of ψ, defined as

$$\operatorname{prox}_{\psi\nu}(t) = \operatorname*{arg\,min}_z \frac{1}{2}\|t - z\|_2^2 + \nu\psi(z),$$

at the point $t = Wx$. Transform sparse coding is often cheap as the proximal mapping of many sparsity penalties can be computed cheaply and in closed form. For instance, when $\psi(z) = \|z\|_0$, then $z = \operatorname{prox}_{\psi\nu}(Wx)$ is computed by setting $z_i = [Wx]_i$ whenever $|[Wx]_i|^2 > \nu^2$, and setting $z_i = 0$ otherwise. This operation is called *hard thresholding*.

Several methods have been proposed to learn a sparsifying transform from data, including algorithms to learn square transforms [41], orthonormal transforms [43], structured transforms [44], and overcomplete transforms consisting of a stack of square transforms [45,46]. Degenerate solutions are prevented by requiring the rows of the learned transform to constitute a well-conditioned frame. In the square case, the transform learning problem can be written

$$\min_{W,Z} \frac{1}{2}\|WX - Z\|_F^2 + \psi(Z) + \frac{1}{2}\|W\|_F^2 - \mu \log|\det W|, \tag{2.2}$$

where X is a matrix whose columns contain training signals and ψ is a sparsity-promoting functional. The first term ensures that the transformed data, WX, is close to the matrix Z, while the second term ensures that Z is sparse. The remaining terms ensure that W is full rank and well-conditioned [41]. Square sparsifying transforms have demonstrated excellent performance in image denoising, magnetic resonance imaging, and computed tomographic reconstruction [47–50].

2.2.3 Analysis Sparsity

Closely related to transform sparsity is the *analysis model.* A signal $x \in \mathbb{R}^N$ satisfies the analysis model if there is a matrix $\Omega \in \mathbb{R}^{K\times N}$, called an analysis operator, such that $\Omega x = z$ is sparse. The analysis model follows by restricting $\eta = \mathbf{0}_K$ in the transform sparsity model.

A typical analysis operator learning algorithm is of the form

$$\min_{\Omega} \psi(\Omega X) + J(\Omega), \tag{2.3}$$

where X are training signals, ψ is a sparsity promoting functional, and J is a regularizer to ensure the learned Ω is informative. In the Analysis K-SVD algorithm, the rows of Ω are constrained to have unit norm, but frame constraints are the most common [51]. Yaghoobi *et al.* observed that learning an analysis operator with $q > K$ rows while using a tight frame constraint resulted in operators consisting of a full rank matrix appended with $q - K$ uniformly zero rows. They instead proposed a uniformly-normalized tight frame (UNTF) constraint, wherein the rows of Ω have equal ℓ_2 norm and constitute a tight frame [52–54].

Hawe *et al.* utilized a similar set of constraints in their GeOmetric Analysis operator Learning (GOAL) framework [55]. They constrained the learned Ω to the set of full column rank matrices with unit-norm rows and solved the optimization problem using a manifold descent algorithm.

Transform and analysis sparsity are closely linked. Indeed, using a variable splitting approach (*e.g.* $Z = \Omega X$) to solve (2.3) leads to algorithms that are similar to transform learning algorithms

[52–54]. The relationships between the transform model, analysis model, and noisy variations of the analysis model have been explored [41]. We focus on the transform model because the proximal interpretation of sparse coding fits nicely within a filter bank interpretation (see Section 2.3.8).

2.2.4 From Patch-Based to Image-Based Models

A link between patch-based and image-based models can be made using the Field of Experts (FoE) model proposed by Roth and Black [56]. They modeled the prior probability of an image as a Markov Random Field (MRF) with overlapping "cliques" of pixels that serve as image patches. Using the so-called product of experts framework, a model for the prior probability of an image patch is expressed as a sparsity-inducing potential function applied to the inner products between multiple "filters" and the image patch. A prior for the entire image is formed by taking the product of the prior for each patch and normalizing.

Continuing in this direction, Chen *et al.* proposed a method to learn an image-based analysis operator using the FoE framework using a bi-level optimization formulation [57]. This approach was recently extended into an iterated filter bank structure called a Trainable Nonlinear Reaction Diffusion (TNRD) network [58]. Each stage of the TNRD network consists of a set of analysis filters, a channelwise nonlinearity, the adjoint filters, and a direct feed-forward path. The filters, nonlinearity, and feed-forward mixing weights are trained in a supervised fashion. The TNRD approach has demonstrated state of the art performance on image denoising tasks.

The TNRD and FoE algorithms are supervised and use the filter bank structure only as a computational tool. In contrast, our approach is unsupervised and uses the theory of perfect reconstruction filter banks to regularize the learning problem.

Cai *et al.* developed an analysis operator learning method based on a filter bank interpretation of the operator [59]. The operator can be thought to act on images, rather than patches. Their approach is fundamentally the same as learning a square, orthonormal, patch-based sparsifying transform [43]. In contrast, our approach does not have these restrictions: we learn a filter bank that is a frame for images, and corresponds to a tall, fat, or square patch-based transform.

These methods fall under the analysis paradigm. In Section 2.3 we show that patch-based analysis models naturally induce a image-based model. In contrast, synthesis patch-based models do not directly lead to an image based model. Figueiredo studied this dichotomy between patch-based synthesis and analysis priors and proposed a method for image-based denoising using patch-based synthesis methods [60].

Image-based modeling using synthesis sparsity can be implemented in an entirely different manner by imposing shift-invariance properties on the synthesis dictionary [61–64]. Briefly, the goal of Convolutional Dictionary Learning (CDL) is to find a set of filters, $\{d_i\}_{i=1}^{N_c}$, such that the training signals can be modeled as $y = \sum_{i=1}^{N_c} d_i * a_i$, where the a_i are sparse. Here, y is an image, not a patch. The filters d_i are required to have compact support so as to limit the number of free parameters in the learning problem. The desired convolutional structure can be imposed by writing the convolution in the frequency domain, but care must be taken to ensure that the d_i remain compactly supported. For further details, see the recent reviews [63, 64].

Finally, Muramatsu *et al.* proposed an approach for the design of multidimensional, multirate, nonseparable, overlapped linear phase perfect reconstruction synthesis filter banks [65–67]. We will refer to a dictionary designed in this manner as a (synthesis) Non-Separable Oversampled Lapped Transforms (NSOLT). Despite using the synthesis sparsity model, the NSOLT design problem shares more in common with our proposed filter bank sparsifying transform learning than the usual CDL problem. We further discuss the NSOLT structure in Section 2.3.4 after the language of polyphase matrices has been established. Differences between NSOLT and our proposed method are discussed in Section 2.4.4.

In the next section, we show that patch-based analysis and transform models, in contrast to synthesis models, are naturally endowed with a convolutional structure.

2.3 From Patch-Based Transforms to Filter Banks

In this section, we illustrate the connections between patch-based sparsifying transforms and multirate finite impulse response (FIR) filter banks. The link between patch-based analysis methods and convolution has been previously established, but used only as a computational tool [56, 57, 59, 68, 69]. Our goal is to illustrate how and when the boundary conditions, patch stride, and a patch-based sparsifying transform combine to form a frame over the space of images.

2.3.1 Frames, Patches, and Images

A set of vectors $\{\omega_i\}_{i=1}^{M}$ in $\mathbb{R}^m$ is a *frame* for $\mathbb{R}^m$ if there exists $0 < A \leq B < \infty$ such that

$$A\|y\|_2^2 \leq \sum_{j=1}^{M} \left|\langle y, \omega_i \rangle\right|^2 \leq B\|y\|_2^2$$

for all $y \in \mathbb{R}^m$ [28]. Equivalently, the matrix $\Omega \in \mathbb{R}^{M \times m}$, with i-th row given by ω_i, is left invertible. The frame bounds A and B correspond to the smallest and largest eigenvalues of $\Omega^T \Omega$, respectively. The frame is *tight* if $A = B$, and in this case $\Omega^T \Omega = A I_n$. The condition number of the frame is the ratio of the frame bounds, B/A. The ω_i are called *frame vectors*, and the matrix Ω implements a *frame expansion*.

Consider a patch-based model using $K \times K$ (vectorized) patches from an $N \times N$ image. We call $\mathbb{R}^{K^2}$ the *space of patches* and $\mathbb{R}^{N \times N}$ the *space of images*. In this setting, transform learning algorithms find a $W \in \mathbb{R}^{N_c \times K^2}$ with rows that form a frame for the space of patches [41, 43–46].

We can extend this W to a frame over the space of images as follows. Suppose the rows of W form a frame with frame bounds $0 < A \leq B$. Let $\mathcal{R}_j : \mathbb{R}^{N \times N} \to \mathbb{R}^{K^2}$ be the linear operator that extracts and vectorizes the j-th patch from the image, and suppose there are M such patches. So long as each pixel in the image is contained in at least one patch, we have

$$\|x\|_F^2 \leq \sum_{j=1}^{M} \|\mathcal{R}_j x\|_2^2 \leq M \|x\|_F^2$$

for all $x \in \mathbb{R}^{N \times N}$. Letting $w^i = W_{i,:}$ denote the i-th row of W, we have for all $x \in \mathbb{R}^{N \times N}$

$$\sum_{j=1}^{M} \sum_{i=1}^{N_c} \left| \langle w^i, \mathcal{R}_j x \rangle \right|^2 \geq \sum_{j=1}^{M} A \|\mathcal{R}_j x\|_2^2 \geq A \|x\|_F^2,$$

$$\sum_{j=1}^{M} \sum_{i=1}^{N_c} \left| \langle w^i, \mathcal{R}_j x \rangle \right|^2 \leq B \sum_{j=1}^{M} \|\mathcal{R}_j x\|_2^2 \leq MB \|x\|_F^2.$$

Because $\langle w^i, \mathcal{R}_j x \rangle_{\mathbb{R}^{K^2}} = \langle \mathcal{R}_j^* w^i, x \rangle_{\mathbb{R}^{N \times N}}$, it follows that the collection $\left\{ \mathcal{R}_j^* w^i \right\}_{i=1, j=1}^{N_c, M}$ forms a frame for the space of images with bounds $0 < A \leq MB$. Thus, every frame over the spaces of patches corresponds to a frame over the space of images. Next, our goal is to determine when the patch extraction operators and the transform, W, form a frame for the space of images but *not* the space of patches.

2.3.2 Patch-Based Sparsifying Transforms as Filter Banks

We consider two ways to represent applying the transform $W \in \mathbb{R}^{N_c \times K^2}$ to the image $x \in \mathbb{R}^{N \times N}$. The usual approach is to form the *patch matrix* $X \in \mathbb{R}^{K^2 \times M^2}$ with j-th column $\mathcal{R}_j x$, as illustrated in Fig. 2.1a. We call the spacing between adjacent extracted patches the *stride* and denote it by s. The extracted patches overlap when $s < K$ and are disjoint otherwise. We assume the stride is the same in both horizontal and vertical directions and evenly divides N. The number of patches,

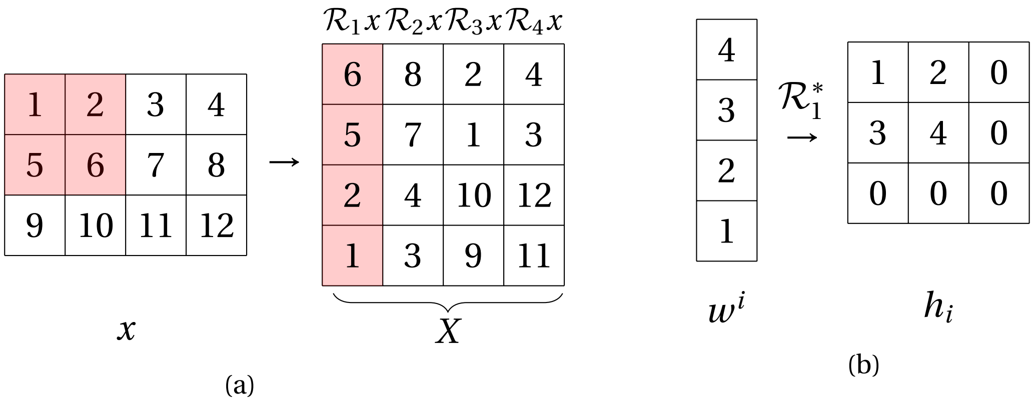


Figure 2.1: (a) Construction of the patch matrix $X \in \mathbb{R}^{4\times4}$ from 2×2 patches of $x \in \mathbb{R}^{3\times4}$ using periodic boundary conditions and a stride of 2. Note that the vectorized patch is "flipped" from the natural ordering; *i.e.*, the top-left pixel in the patch is the final element of the vector. (b) Obtaining the impulse response h_i from w^i.

M^2, depends on the boundary conditions and patch stride; *e.g.* $M^2 = N^2/s^2$ if periodic boundary conditions are used. The patch matrix for the transformed image is $WX \in \mathbb{R}^{N_c \times M^2}$.

Our second approach eliminates patches and their vectorization by viewing WX as the output of a multirate filter bank with 2D FIR filters and input x. Let

$$\mathcal{H} : \mathbb{R}^{N\times N} \to \mathbb{R}^{N_c} \otimes \mathbb{R}^{M\times M}$$

be this filter bank operator, which transforms an $N \times N$ image into a three-dimensional array formed as a stack of N_c output images, each of size $M \times M$.

We build $\mathcal{H}$ from a collection of downsampled convolution operators. For $i = 1, 2, \ldots N_c$, we define the i-th channel operator $\mathcal{H}_i : \mathbb{R}^{N\times N} \to \mathbb{R}^{M\times M}$ such that $[\mathcal{H}_i x]_{a,b} = [h_i * x]_{sa,sb}$. The stride s dictates the downsampling level, and the patch extraction boundary conditions determine the convolution boundary conditions; in particular, if periodic boundary conditions are used, then $\mathcal{H}_i$ implements cyclic convolution. The impulse response h_i is obtained from the i-th row of W as $\mathcal{R}_1^* w^i$. This matrix consists of a $K \times K$ submatrix embedded into the upper-left corner of an $N \times N$ matrix of zeros as illustrated in Fig. 2.1b. [1]

Finally, we construct $\mathcal{H}$ by "stacking" the channel operators: $\mathcal{H} = \sum_{i=1}^{N_c} e_i \otimes \mathcal{H}_i$, where e_i is the i-th standard basis vector in $\mathbb{R}^{N_c}$ and $\otimes$ denotes the Kronecker (or tensor) product. With this definition, $y = \mathcal{H}x = \sum e_i \otimes \mathcal{H}_i x = \sum e_i \otimes y_i$. The filter bank structure is illustrated in Fig. 2.2. We refer to $\mathcal{H}$ constructed in this form as a *filter bank sparsifying transform*. The following

[1] In the case of cyclic convolution, $\mathcal{R}_1^* w^i$ is exactly the impulse response of the i-th channel, but only the non-zero portion of $\mathcal{R}_1^* w^i$ is the impulse response when using linear convolution. In a slight abuse of terminology, we call $\mathcal{R}_1^* w^i$ the impulse response in both instances.

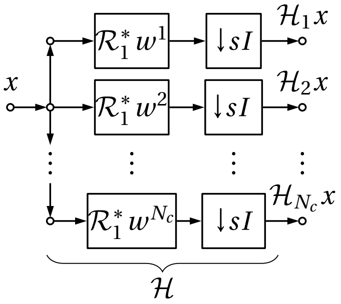


Figure 2.2: Analysis filter bank $\mathcal{H}$ generated by a sparsifying transform W and stride length s.

proposition links WX and $\mathcal{H}x$.

Proposition 2.1. *Let $X \in \mathbb{R}^{K^2 \times M^2}$ be a patch matrix for image X, and let $W \in \mathbb{R}^{N_c \times K^2}$. The rows of WX can obtained by passing x through the N_c channel, 2D FIR multirate analysis filter bank $\mathcal{H}$ and vectorizing the channel outputs.*

A proof for 1D signals is given in Appendix A.1. The proof for 2D is similar, using vector indices.

Proposition 2.1 connects the local, patch-extraction process and the matrix W to a filter bank operator that acts on images. Unlike convolutional synthesis models, patch-based analysis operators naturally have a convolutional structure.

Next, we investigate connections between the frame properties of $\mathcal{H}$ and the combination of W and the patch extraction scheme. Our primary tool is the polyphase representation of filter banks [7, 8]. Consider the image x as a 2D sequence $x[n_1, n_2]$ for $0 \le n_1, n_2 \le N-1$. The z-transform of the (a,b)-th polyphase component of x is z-transform $\hat{X}_{a,b}(\mathbf{z}) = \sum_{n_1,n_2} x[n_1 \cdot s + a, n_2 \cdot s + b] z_1^{-n_1} z_2^{-n_2}$ of the shifted and downsampled sequence, where $\mathbf{z} = [z_1, z_2] \in \mathbb{C}^2$ and $0 \le a, b \le s-1$. The polyphase representation for the sequence x is formed by stacking the polyphase components in lexicographical order into a single $\hat{X}(\mathbf{z}) = \left[X_{0,0}(\mathbf{z}), \ldots, X_{s-1,s-1}(\mathbf{z})\right]^T \in \mathbb{C}^{s^2}$.

The filter bank $\mathcal{H}$ has a polyphase matrix $\hat{H}(\mathbf{z}) \in \mathbb{C}^{Nc \times s^2}$ formed by stacking the polyphase representations of each channel into a row, and stacking the N_c rows. Explicitly,

$$\hat{H}(\mathbf{z}) = \begin{bmatrix} \hat{H}^0_{0,0}(\mathbf{z}) & \hat{H}^0_{0,1}(\mathbf{z}) & \ldots & \hat{H}^0_{s,s}(\mathbf{z}) \\ \hat{H}^1_{0,0}(\mathbf{z}) & \hat{H}^1_{0,1}(\mathbf{z}) & \ldots & \hat{H}^1_{s,s}(\mathbf{z}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{H}^{N_c-1}_{0,0}(\mathbf{z}) & \hat{H}^{N_c-1}_{0,1}(\mathbf{z}) & \ldots & \hat{H}^{N_c-1}_{s,s}(\mathbf{z}) \end{bmatrix},$$

where $\hat{H}^i_{a,b}(\mathbf{z})$ is the (a,b)-th polyphase component of the i-th filter in $\mathcal{H}$. The entries of $\hat{H}(\mathbf{z})$ are, in general, bi-variate polynomials in $\mathbf{z} = [z_1, z_2]$. The output of the filter bank, $y = \mathcal{H}x$, can be written in the polyphase domain as $\hat{Y}(\mathbf{z}) = \hat{H}(\mathbf{z})\hat{X}(\mathbf{z})$, where the i-th element of the vector $\hat{Y}(\mathbf{z})$ is the z-transform of the i-th output channel.

Many important properties of $\mathcal{H}$ are tied to its polyphase matrix. An analysis filter bank $\mathcal{H}$ is said to be *perfect reconstruction* (PR) if there is a (synthesis) filter bank $\mathcal{G}$ such that $\mathcal{G}\mathcal{H} = \mathcal{I}$, or in the polyphase domain $\hat{G}(\mathbf{z})\hat{H}(\mathbf{z}) = I$. A filter bank is PR if and only if $\hat{H}(\mathbf{z})$ is full column rank on the unit circle [12]. A filter bank is said to be *orthonormal* if $\mathcal{H}^*\mathcal{H} = \mathcal{I}$, that is, the filter bank with analysis section $\mathcal{H}$ and synthesis section $\mathcal{H}^*$ is an identity mapping on $\mathbb{R}^{N\times N}$. In the polyphase domain, this corresponds to

$$\hat{H}^*(\mathbf{z}^{-1})\hat{H}(\mathbf{z}) = I, \tag{2.4}$$

where the star superscript denotes Hermitian transpose and $\mathbf{z}^{-1} = [z_1^{-1}, z_2^{-1}]$ [7, 29]. A matrix satisfying (2.4) is *paraunitary*, and $\hat{H}^*(\mathbf{z}^{-1})$ is the *paraconjugate* of $\hat{H}(\mathbf{z})$.

A PR filter bank implements a frame expansion over the space of images, and an orthonormal filter bank implements a tight frame expansion over the same space [27, 70]. The frame vectors are the collection of the shifts of the impulse responses of each channel, and are precisely the collection $\left\{\mathcal{R}_j^* w^i\right\}$ discussed in Section 2.3.1. The link between patch-based transforms and filter banks does not directly lead to new transform learning algorithms, as the characterization and construction of multidimensional PR filter banks is hard due to the lack of a multidimensional spectral factorization theorem [8–11].

Next, we study we illustrate the connections between patch-based sparsifying transforms and perfect reconstruction filter banks as a function of the stride length. We show that in certain cases the PR condition takes on a simple form.

2.3.3 Perfect Recovery: Non-Overlapping Patches

Consider $s = K$, so that the extracted patches do not overlap. Applying the sparsifying transform W to non-overlapping patches is an instance of a *block transform* [71]. Block transforms are found throughout in signal processing applications; for example, the JPEG compression algorithm. Block transforms are viewed as a decimated FIR filter bank with uniform downsampling by K in each dimension, consistent with Proposition 2.1.

It is informative to view patch-based transform learning algorithms through the lens of block transformations. Because we downsample by K in each dimension, and the filters are of size $K \times K$, the polyphase matrix $\hat{H}(\mathbf{z})$ is constant in (independent of) $\mathbf{z}$ and is equal to W. This gives a direct connection between the PR properties of $\mathcal{H}$, which acts on images, and W, which acts on patches. Patch-based transform learning algorithms enforce either invertibility of W (in the square case) or invertibility of $W^T W$ (in the overcomplete case), and thus $\mathcal{H}$ is PR. If W is orthonormal, so too is $\mathcal{H}$.

2.3.4 Perfect Recovery: Partially Overlapping Patches

Next, consider patches extracted using a stride $1 < s < K$. While WX is no longer a block transformation, it is related to a *lapped transformation* [71]. Lapped transforms aim to reduce artifacts that arise from processing each block (patch) independently by allowing for overlap between neighboring blocks. Many lapped transforms, such as the lapped orthogonal transform, the extended lapped transform, and the generalized lapped orthogonal transform [72], enjoy both the PR property and efficient implementation.

Lapped transforms were designed for signal coding applications. The number of channels in the filter bank is decreased as the degree of overlap increases, so that the number of transform coefficients using a lapped transform is the same as using a non-lapped transform. While redundancy may be undesirable in certain coding applications, it aids the solution of inverse problems by allowing for richer and more robust signal models [73]. We allow the stride length to decrease while keeping the number of channels fixed, and interpret WX as a "generalized" lapped transform. When the stride is less than K, W no longer corresponds to the polyphase matrix of the filter bank $\mathcal{H}$; instead, the polyphase matrix $\hat{H}(\mathbf{z})$ will contain high-order, 2D polynomials. While the filter bank may still be PR, the PR property is not directly implied by invertibility of W.

We can learn a PR generalized lapped transform by enforcing the more restrictive PR conditions for non-overlapping patches, that is, invertibilty of $W^T W$. When $s = 1$, this technique is equivalent to cycle spinning, which was developed to add shift-invariance to decimated wavelet transforms [74]. When $1 < s < K$, we can interpret $\mathcal{H}x$ as cycle spinning without all possible shifts.

Muramatsu *et al.* proposed a different method to learn a PR *synthesis* non-separable oversampled lapped transform (NSOLT) [65–67]. The filter bank is designed such that each channel consists of linear phase filters; thus each channel consists of either symmetric or anti-symmetric filters.

Further, the filter bank is parameterized by a certain lattice structure that implicitly ensures the NSOLT implements a tight-frame expansion and is thus PR. This lattice structure leads to a particular factorization of the (paraunitary) polyphase matrix; in two dimensions, we have $\hat{\mathbf{H}}(\mathbf{z}) = \mathbf{G}_1(z_1)\mathbf{G}_2(z_2)\mathbf{H}_0$, where $\mathbf{G}_i(z_i)$ is a univariate polynomial matrix of specified order and $\mathbf{H}_0$ is constant in $\mathbf{z}$. Each of these matrices is further parameterized to lead to a tractable optimization problem; see [67] for details.

Patch-based sparsifying transforms and NSOLTs are both parameterized in terms of a polyphase matrix, and thus lead to filter banks with compactly supported filters. This is in stark contrast to the usual convolutional dictionary learning problems, where variable splitting methods are often used to obtain both a convolutional structure and compactly supported filters.

2.3.5 Perfect Recovery: Maximally Overlapping Patches

Finally, consider extracting maximally overlapping patches by setting $s = 1$. The resulting filter bank $\mathcal{H}$ is undecimated and the Gram operator $\mathcal{H}^*\mathcal{H}$ is shift invariant. As there is no downsampling, the polyphase representations of x and y are the z-transforms of the sequences x and y. The polyphase matrix of $\mathcal{H}$ is the column vector $\hat{H}(\mathbf{z}) = [\hat{H}_1(\mathbf{z}), \dots \hat{H}_{N_c}(\mathbf{z})]^T$ where $\hat{H}_i(\mathbf{z})$ is the z-transform of $h_i = \mathcal{R}_1^* w^i$.

An undecimated linear convolution filter bank is PR if and only if its filters have no common zeros on the unit circle; *i.e.*, each frequency must pass through at least one channel of the filter bank [12]. When evaluated on the unit circle the z-transform becomes the Discrete Time Fourier Transform (DTFT), defined for $h \in \mathbb{R}^{K \times K}$ as

$$\hat{H}(\boldsymbol{\omega}) = \sum_{n_1=0}^{K-1} \sum_{n_2=0}^{K-1} h[n_1, n_2] e^{-j\omega_1 n_1} e^{-j\omega_2 n_2},$$

where $\boldsymbol{\omega} = [\omega_1, \omega_2]$ with $\omega_1, \omega_2 \in [0, 2\pi)$. Now, the polyphase matrix is full rank on the unit circle if and only if

$$\varphi(\boldsymbol{\omega}) \triangleq \sum_{i=1}^{N_c} \left| \hat{H}_i(\boldsymbol{\omega}) \right|^2 > 0 \quad \forall w_1, w_2 \in [0, 2\pi), \tag{2.5}$$

where $\varphi(\boldsymbol{\omega})$ is the DTFT of the impulse response of $\mathcal{H}^*\mathcal{H}$ and is an even, real, non-negative, 2D trigonometric polynomial with maximum component order $K - 1$. Explicitly,

$$\varphi(\boldsymbol{\omega}) = \sum_{n_1=-K+1}^{K-1} \sum_{n_2=-K+1}^{K-1} \tilde{h}[n_1, n_2] \cos(\omega_1 n_1) \cos(\omega_2 n_2),$$

where the impulse response of $\tilde{h}$ is $\mathcal{H}^*\mathcal{H}$ is the sum of the channel-wise autocorrelations; that is,

$$\tilde{h}[n_1, n_2] = \sum_{i=1}^{N_c} \sum_{l_1=-\infty}^{\infty} \sum_{l_2=-\infty}^{\infty} h_i[l_1, l_2] h_i[l_1 - n_1, l_2 - n_2].$$

Direct verification of the PR condition (2.5) is NP-hard for $K \geq 2$, underlining the difficulty of multidimensional filter bank design [3, 4]. We sidestep the difficulty of working with (2.5) by developing the PR condition when image patches are extracted using periodic boundary conditions. The resulting filter bank implements *cyclic* convolution. Afterwards, we show that under certain conditions, the PR property of a cyclic convolution filter bank implies the PR property of a linear convolution filter bank constructed from the same filters.

2.3.6 Periodic Boundary Conditions / Cyclic Convolution

If image patches are extracted using periodic boundary conditions, the channel operators $\mathcal{H}_i : \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$ implement cyclic convolution and are diagonalized by the 2D Discrete Fourier Transform (DFT). Let $\mathcal{F}$ be the orthonormal 2D-DFT operator such that

$$(\mathcal{F}h_i)[\mathbf{k}] = N^{-1} \sum_{n_1=0}^{K-1} \sum_{n_2=0}^{K-1} h_i[n_1, n_2] e^{-j\frac{2\pi k_1 n_1}{N}} e^{-j\frac{2\pi k_2 n_2}{N}}$$

for $\mathbf{k} = [k_1, k_2]$ and $0 \le k_1, k_2 < N$; that is, the length N 2D-DFT of the filter h_i padded with $N-K$ zeros in each dimension. Define $\mathcal{D}_i \in \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ as the operator that multiplies pointwise by $\mathcal{F}h_i$: for $u \in \mathbb{C}^{N\times N}$, we have $(\mathcal{D}_i u)(\mathbf{k}) = (\mathcal{F}h_i)(\mathbf{k}) \cdot u(\mathbf{k})$. The cyclic convolution operator $\mathcal{H}_i$ has eigenvalue decomposition $\mathcal{F}^*\mathcal{D}_i\mathcal{F}$. We can use this channel-wise decomposition to find the spectrum of $\mathcal{H}^*\mathcal{H}$.

Lemma 2.1. *The N^2 eigenvalues of the undecimated cyclic analysis-synthesis filter bank $\mathcal{H}^*\mathcal{H}$ are given by $\sum_{i=1}^{N_c} |(\mathcal{F}h_i)[\mathbf{k}]|^2$ for $\mathbf{k} = [k_1, k_2]$ and $0 \le k_1, k_2 < N$.*

Proof. We have

$$\begin{aligned}\mathcal{H}^*\mathcal{H} &= \sum_{i=1}^{N_c} (e_i \otimes \mathcal{H}_i)^* (e_i \otimes \mathcal{H}_i) = \sum_{i=1}^{N_c} \mathcal{H}_i^* \mathcal{H}_i \\ &= \mathcal{F}^* \left(\sum_{i=1}^{N_c} \mathcal{D}_i^* \mathcal{D}_i \mathcal{F} \right) = \mathcal{F}^* \mathcal{D} \mathcal{F},\end{aligned}$$

where $(\mathcal{D}u)[\mathbf{k}] = \sum_{i=1}^{N_c} |(\mathcal{F}h_i)[\mathbf{k}]|^2 \cdot u[\mathbf{k}]$. □

The quantity $|(\mathcal{F}h_i)[\mathbf{k}|]^2$ is the squared magnitude response of the i-th filter evaluated at the DFT frequency $\mathbf{k}$, and the eigenvalues of $\mathcal{H}^*\mathcal{H}$ are the sum over the N_c channels of these squared magnitude responses. As the DFT consists of samples of the DTFT, by Lemma 2.1 and (2.5), the eigenvalues of $\mathcal{H}^*\mathcal{H}$ can be seen to be samples of the trigonometric polynomial $\varphi(\boldsymbol{\omega})$ over the set $\Theta_N = \left\{ \left(\frac{2\pi k_1}{N}, \frac{2\pi k_2}{N} \right) : 0 \le k_1, k_2 < N \right\}$.

Recall that $\mathcal{H}$ implements a frame expansion only if the smallest eigenvalue of $\mathcal{H}^*\mathcal{H}$ is strictly positive [28]. We have the following PR condition for cyclic convolution filter banks.

Corollary 2.2. *The undecimated cyclic filter bank $\mathcal{H}$ implements a frame expansion for $\mathbb{R}^{N\times N}$ if and only if $\sum_{i=1}^{N_c} |(\mathcal{F}h_i)[\mathbf{k}]|^2 > 0$ for $0 \leq k_1, k_2 < N$. If $\mathcal{H}$ implements a frame expansion, the upper and lower frame bounds are $\min_{\mathbf{k}} \sum_{i=1}^{N_c} |(\mathcal{F}h_i)[\mathbf{k}]|^2$ and $\max_{\mathbf{k}} \sum_{i=1}^{N_c} |(\mathcal{F}h_i)[\mathbf{k}]|^2$.*

Whereas the PR condition for a linear convolution filter bank must hold over the unit circle, the PR condition for cyclic convolution filter bank involves only the N^2 DFT frequencies.

The factorization $\mathcal{H}^*\mathcal{H} = \mathcal{F}^*\mathcal{D}\mathcal{F}$ also provides an easy way to compute the (minimum norm) synthesis filter bank $\mathcal{H}^\dagger$ that satisfies $\mathcal{H}^\dagger\mathcal{H} = \mathcal{I}$. We have $\mathcal{H}^\dagger = (\mathcal{H}^*\mathcal{H})^{-1}\mathcal{H}^*$, and the necessary inverse is given by $(\mathcal{H}^*\mathcal{H})^{-1} = \mathcal{F}^*\mathcal{D}^{-1}\mathcal{F}$.

2.3.7 Return to Linear Convolution

We now want to link the PR conditions for cyclic and linear convolution filter banks. The inequalities of Chapter 1 provide the tool we need. Recall that we showed that the minimum value of a real, multivariate trigonometric polynomial can be lower bounded given sufficiently many uniformly spaced samples of the polynomial, provided that the polynomial does not vary too much over the sampling points.

With $\varphi(\boldsymbol{\omega})$ is defined by (2.5), then κ_N is the frame condition number of a cyclic convolution filter bank operating on $N \times N$ images.

Corollary 1.3 is the link between PR properties of cyclic and linear convolution filter banks we desired, and we have the following PR condition for linear convolution filter banks.

Corollary 2.3. *Let $\mathcal{H}_C$ be an undecimated cyclic convolution filter bank with $K \times K$ filters that operates on $N \times N$ images, with frame condition number κ_N. Let $\mathcal{H}$ be a linear convolution filter bank constructed from the same filters as $\mathcal{H}_C$. Then $\mathcal{H}$ is PR if $\kappa_N \leq \frac{N}{K-1} - 1$.*

Proof. This follows immediately from Corollary 1.3. Set $d = 2$ and $n = K - 1$ in (1.14). □

Corollary 2.3 states that well-conditioned PR cyclic convolution filter banks, with filters that are short relative to image size N, are also PR *linear* convolution filter banks.

The PR conditions of Corollaries 2.2 and 2.3 are significantly more general than the patch-based PR conditions. For example, $W \in \mathbb{R}^{N_c \times K^2}$ can be left-invertible only if $N_c \geq K^2$. The PR conditions of Corollaries 2.2 and 2.3 have no such requirements; indeed, a single channel "filter bank" can be PR. Our PR conditions are easy to check, requiring only the 2D DFT of N_c small filters.

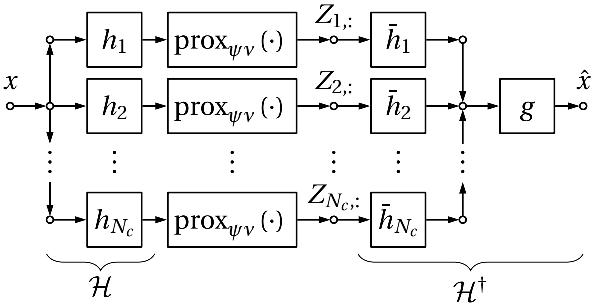


Figure 2.3: Analysis-synthesis filter bank generated by sparsifying transform W and separable sparsity penalty ψ. Here, $h_i = \mathcal{R}_1^* w^i$, the impulse response $\bar{h}_i$ is the flipped version of h_i, and g is the impulse response for the filter $(\mathcal{H}^*\mathcal{H})^{-1}$.

2.3.8 The Role of Sparsification

We have interpreted the transformed image patches WX as the output of a filter bank. The sparse matrix Z in (2.1) can be viewed as passing the filter bank output through a nonlinear function implementing $\mathrm{prox}_{\psi\nu}(WX)$. This interpretation is particularly appealing whenever ψ is coordinate-wise separable, meaning $\psi(z) = \sum_i \psi(z_i)$. Then the transform sparse code for the j-th channel depends only on the j-th filtered channel and is given by $\mathrm{prox}_{\psi\nu}\left(\mathcal{H}_j x\right)$. The resulting nonlinear analysis-synthesis filter bank is illustrated in Fig. 2.3. If the input signal x is indeed perfectly sparsifiable by the filter bank (*i.e.*, $\mathcal{H}x = \mathrm{prox}_{\psi\nu}(\mathcal{H}x)$), then the output of the analysis stage is invariant to the application of the nonlinearity and the entire system functions as an identity operator.

We can replace the usual soft or hard thresholding functions by exotic nonlinearities, such as the firm thresholding function [75] or generalized p-shrinkage functions [76]. These nonlinearities have led to marginally improved performance in image denoising [57] and compressed sensing [77]. Alternatively, we can abandon the interpretation of the nonlinearity as a proximal operator and instead learn a custom nonlinearity for each channel, either in a supervised setting [58, 78] or in an unsupervised setting with a Gaussian noise model [79].

Filter bank sparsifying transforms share many similarities with convolutional autoencoder (CAE) [80]. Both consist of a filter bank followed by a channelwise nonlinearity. However, in the case of a CAE, the "decoder" $\mathcal{H}^\dagger$ is typically discarded and the output of the "encoder", $\mathrm{prox}_{\psi\nu}(\mathcal{H}x)$, is passed into additional layers for further processing.

2.3.9 Principal Component Filter Banks

The previous sections have shown that transform learning can be viewed as adapting a filter bank to sparsify our data. A similar problem is the design of *principal component filter banks* (PCFB). Let $\mathscr{C}$ denote a set of orthonormal filter banks, such as N_c-channel filter banks with downsampling matrix M, and let x be a given input signal. A filter bank $\mathcal{H}^{\mathscr{P}}$ is said to be a PCFB for the class $\mathscr{C}$ and the input x if, for any $\mathcal{H} \in \mathscr{C}$ and all $m = 1, \ldots, N_c$,

$$\sum_{i=1}^{m} a_i^2 \geq \sum_{i=1}^{m} b_i^2 \tag{2.6}$$

where a_i and b_i are the ℓ_2 norms of the i-th channel of $\mathcal{H}^{\mathscr{P}} x$ and $\mathcal{H}x$, respectively [81]. A PCFB provides compaction of the output energy into lower-indexed channels, and thus minimal ℓ_2 error when reconstructing a signal from $m < M$ filtered components. The existence and design of PCFBs in 1D is well studied [82]. However, the design of multidimensional FIR PCFBs is again made difficult due to the lack of a multidimensional spectral factorization theorem, although suboptimal algorithms exist [83, 84].

There are superficial similarities between the design of PCFBs and transform learning, especially when W is restricted to be square and orthonormal. The sparsity of the transformed signal implies a form of energy compaction. However, we impose no constraints on location of non-zero coefficients and thus the learned transform is unlikely to satisfy the majorization property (2.6). Further, an orthogonal W matrix induces an orthogonal filter bank only if non-overlapping patches are used. The PCFB for such a block transformation is known to be the Karhunen-Loeve transformation of the data [85], from which the learned transform can differ substantially [41]. Conversely, the energy majorization property (2.6) does not imply sparsity of the channel outputs, and a PCFB will not, in general, be a filter bank sparsifying transform.

2.4 Learning a Sparsifying Filter Bank

We briefly review methods to incorporate an adaptive sparsity model in the solution of inverse problems. We consider two paradigms: in the "universal" paradigm, our sparsifying transform $\mathcal{H}$ is learned off-line over a set of training data. In the "adaptive" paradigm, the transform is learned during the solution of the inverse problem, typically by alternating between a signal recovery phase and a transform update phase. For synthesis dictionary learning it has been reported that the adaptive method typically works better for lower noise levels while the universal method shines in high noise [42]. In both paradigms, we learn sparsifying transform by minimizing a function that is designed to encourage “good” transforms.

2.4.1 Problem Formulation

We now develop a method to learn an undecimated filter bank sparsifying transform that takes advantage of the flexibility granted by the PR conditions of Corollaries 2.2 and 2.3. Let x be a training signal, possibly drawn from a set of training signals. We wish to learn a filter bank sparsifying transform that satisfies four properties:

(D1) $\mathcal{H}x$ should be sparse

(D2) $\mathcal{H}$ should be left invertible and well conditioned

(D3) $\mathcal{H}$ should contain no duplicated or uniformly zero filters

(D4) $\mathcal{H}$ should have few degrees of freedom

Properties (D1) – (D3) ensure our transform is “good” in that it sparsifies the training data, is a well-behaved frame over the space of images, and is not overly redundant. Property (D4) ensures good generalization properties under the universal paradigm and prevents overfitting under the adaptive paradigm.

As with previous transform learning approaches, we satisfy (D1) by minimizing $\frac{1}{2}\|\mathcal{H}x - z\|_2^2 + \nu\psi(z)$ where ψ is a sparsity-promoting functional. The first of term is called the *sparsification error*, as it measures the distance between $\mathcal{H}x$ and its sparsified version, z.

We satisfy (D4) by writing the action of the sparsifying transform on the image as WX, where $W \in \mathbb{R}^{N_c \times K^2}$ and X is formed by extracting and vectorizing $K \times K$ patches with unit stride. This parameterization ensures that we have the desired filter bank structure, and that the learned filters are compactly supported and have only $N_c K^2$ free parameters.

This is a key difference between convolutional analysis-based methods, such as ours, and synthesis-based convolutional dictionary learning; learning a convolutional dictionary requires careful parameterization to get both the desired convolutional structure and filters of compact support.

We emphasize that WX and $\mathcal{H}x$ are equivalent modulo a reshaping operation. Both expressions should be thought of independently of the computational tool used to calculate the results; WX can be implemented using Fourier-based fast convolution algorithms, just as $\mathcal{H}x$ can be implemented by dense matrix-matrix multiplication. We further elaborate on this point in Section 2.4.3. We choose to write the filter bank application as WX so that we can express the sparsification error directly in terms of W; in particular, we have

$$f(W, Z, x) = \frac{1}{2}\|WX - Z\|_F^2, \tag{2.7}$$

where the j-th row of $Z \in \mathbb{R}^{N_c \times N^2}$ is the sparse code for the j-th channel output. We can learn a transform over several images by summing the sparsification error for each image.

We promote transforms that satisfy (D2) through the penalty $\frac{1}{2}\sum_{j=1}^{N_c}\|W_{j,:}\|_2^2 - \log\det \mathcal{H}^*\mathcal{H}$. The log determinant term ensures that no eigenvalues of $\mathcal{H}^*\mathcal{H}$ become zero, while the ℓ_2 norm term ensures that the filters do not grow too large. This penalty can be written as $\sum_{j=1}^{N^2} \lambda_i - \log \lambda_i$, where the λ_i are the eigenvalues of $\mathcal{H}^*\mathcal{H}$ as given by Lemma 2.1. Our proposed penalty serves the role of the final two terms of the patch-based objective (2.2). The key difference is that the patch-based regularizer acts on the singular values of W, while the proposed regularizer acts on the singular values of $\mathcal{H}$.

To satisfy (D4) we write the eigenvalues λ_i in terms of the matrix W. Let $F \in \mathbb{C}^{N^2 \times N^2}$ denote the matrix that computes the $N \times N$ orthonormal 2D-DFT for a vectorized signal, and let $\bar{F} \in \mathbb{C}^{N^2 \times K^2}$ represent the $N \times N$ 2D-DFT of a zero-padded and vectorized $K \times K$ signal. The i-th column of $\bar{F}W^T$ contains the (vectorized) 2D-DFT of the i-th filter. Then $\lambda_i = \sum_{j=1}^{N_c} \left|\bar{F}W^T\right|_{i,j}^2$, and

$$\log\det \mathcal{H}^*\mathcal{H} = \sum_{i=1}^{N_F^2} \log\left(\sum_{j=1}^{N_c} \left|\bar{F}W^T\right|_{i,j}\right),$$

where the absolute value and squaring operations are taken pointwise. We can reduce the computational and memory burden of the algorithm by using smaller $N_F \times N_F$ DFTs, provided that Corollary 2.3 implies the corresponding linear convolution filter bank is PR. We take $N_F = 4K$, which is suitable for filter banks with condition number less than 3.

Similar to earlier work on analysis operator learning [52–54], we found that our tight frame penalty often resulted in transforms with many uniformly zero filters. We prevent zero-norm filters by adding the log-barrier penalty $\sum_{j=1}^{N_c} - \log\left(\|W_{j,:}\|_2^2\right)$. The combined regularizer is written as

$$J_1(W) = \frac{1}{2}\sum_{i=1}^{N_c}\|W_{i,:}\|_2^2 - \sum_{i=1}^{N_F^2} \log\left(\sum_{j=1}^{N_c} \left|\bar{F}W^T\right|_{i,j}^2\right) - \sum_{i=1}^{N_c} \log\left(\|W_{i,:}\|_2^2\right). \tag{2.8}$$

The following proposition (proved in Appendix A.2) indicates that J_1 promotes filter bank transforms that satisfy (D2).

Proposition 2.2. *Let $W^\sharp$ be a minimizer of J_1, and let $\mathcal{H}$ be the undecimated cyclic convolution* *filter bank generated by the rows of* $W^\sharp$. *Then* $\mathcal{H}$ *implements a uniformly normalized tight frame expansion over the space of images, with filter squared norms equal to* $2(1+N_F^2/N_c)$ *and frame constant* $2(1+N_c/N_F^2)$.

Finally, we would like to discourage learning copies of the same filter. To that end, we define the coherence between rows i and j of W as

$$\Gamma_{i,j}(W) \triangleq \frac{\langle W_{i,:}, W_{j,:}\rangle}{\|W_{i,:}\|_2 \|W_{j,:}\|_2}.$$

One option is to apply a log barrier to the squared coherence between each pair of filters [55]:

$$J_2(W) = \sum_{1\le i<j\le N_c} -\log\Big(1-\big(\Gamma_{i,j}(W)\big)^2\Big). \tag{2.9}$$

This penalty works well whenever the filters have small support ($K \le 8$). For larger filters, we observed the algorithm often learned filters with disjoint support that are shifted versions of one another. These filters do not cause a large value in (2.9), yet provide no advantage over a single filter. We modify our coherence penalty to discourage filters that differ by only a linear phase term by applying (2.9) to the squared magnitude responses of our filters. This coherence penalty naturally promotes zero-mean filters, as the coherence between two non-zero-mean filters can be reduced simply by removing their mean.

Our learning problem is written as

$$\min_{W,Z} f(W,Z,x) + \mu J_1(W) + \lambda J_2(W) + \nu\psi(Z). \tag{2.10}$$

The scalar $\mu > 0$ controls the strength of the UNTF penalty and should be large enough that the learned filter bank is well conditioned, so that approximating the eigenvalues using $N_F \times N_F$ DFTs remains valid. The non-negative scalar parameters λ and ν control the emphasis given to the coherence and sparsity penalties, respectively.

2.4.2 Optimization Algorithm

We use an alternating minimization algorithm to solve (2.10). In the *sparse coding step*, we fix W and solve (2.10) for Z. In the second stage, called the *transform update step*, we update our transform W by minimizing (2.10) with fixed Z. We use superscripts to indicate iteration number, and we take $\mathcal{H}^{(k)}$ to mean the filter bank generated using filters contained in the rows of $W^{(k)}$.

The sparse coding step reduces to

$$Z^{(k+1)} = \operatorname*{argmin}_{Z} \frac{1}{2} \| W^{(k)} X - Z \|_F^2 + \nu \psi(Z)$$

with solution $Z^{(k+1)} = \operatorname{prox}_{\psi\nu}\left(\mathcal{H}^{(k)} x\right)$. Next, with $Z^{(k+1)}$ fixed, we update W by solving

$$W^{(k+1)} = \operatorname*{argmin}_{W} f(W, Z^{(k+1)}, x) + \mu J_1(W) + \lambda J_2(W). \tag{2.11}$$

Unlike the square, patch-based case, we do not have a closed-form solution for (2.11) and we must resort to an iterative method. The limited-memory BFGS (L-BFGS) algorithm works well in practice. The necessary gradients are

$$\nabla_W f(W, Z, x) = 2WXX^T - 2XZ^T, \tag{2.12}$$

$$\nabla_W \log\det \mathcal{H}^* \mathcal{H} = 2W\bar{F}^* \operatorname{ddiag}\left(\left|\bar{F} W^T\right|^2 \mathbf{1}_{N_c}\right)^{-1} \bar{F}, \tag{2.13}$$

$$\frac{\partial}{W_{r,s}} \sum_{i=1}^{N_c} \log\left(\| W_{i,:} \|_2^2\right) = \frac{2 W_{r,s}}{\| W_{r,:} \|_2^2},$$

$$\frac{\partial J_2(W)}{\partial W_{r,s}} = \sum_{i=1, i \neq r}^{N_c} \frac{W_{i,s}[WW^T]_{i,r} - W_{r,s}[WW^T]_{i,r}^2 \| W_{r,:} \|_2^{-2}}{\| W_{i,:} \|_2^2 \cdot \| W_{r,:} \|_2^2 - [WW^T]_{i,r}^2}.$$

2.4.3 Computational Considerations

The primary bottleneck in using L-BFGS to solve (2.11) is the line search step, which requires multiple evaluations of the objective function (2.10) with fixed Z. The cost of this computation is dominated by evaluation of (2.7). With X and Z fixed, we precompute and store the small matrices $G = XX^T$ and $Y = XZ^T$. The sparsification residual is evaluated as

$$\operatorname{trace}\left(W^T W G\right) - 2 \cdot \operatorname{trace}(WY) + \| Z \|_F^2$$

and requires only small matrix-matrix products.

In the patch-based case, evaluating WX using dense matrix multiplication requires $\mathcal{O}(N_c K^2 N^2)$ floating point operations (FLOPS). The filter bank structure of $\mathcal{H}$ naturally leads to efficient calculation of $\mathcal{H}x$ through the use of Fourier-based convolution methods. For simplicity, we will restrict our attention to radix-2 FFT algorithms and assume that both K and N are powers of 2.

The usual Fourier-based circular convolution methods require adding zeros until both signals are of the same size. In our case, we must extend each row of W to be of length N. Passing the

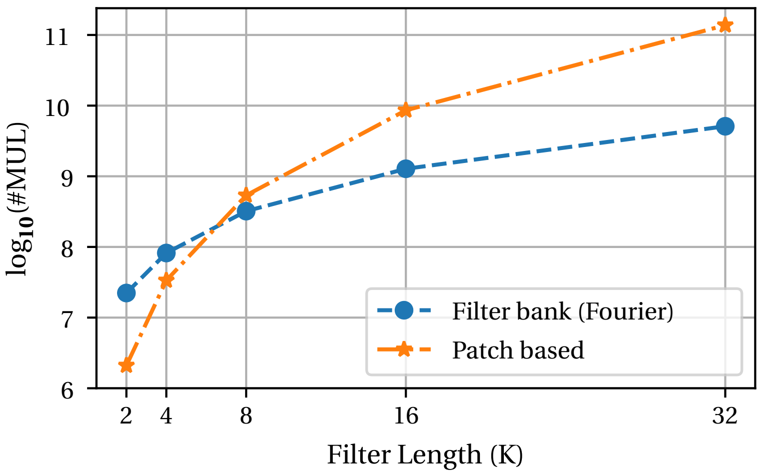


Figure 2.4: Number of multiplications needed to apply a square filter bank ($N_c = K^2$) to a 512 × 512 image using Fourier and patch-based methods.

input signal through a single filter will require $\mathcal{O}(N_c N^2 \log N)$ FLOPS, representing a $K^2/\log(N)$ reduction over the patch-based case. For the typical sizes of $N = 512$ and $K = 8$, this is roughly a factor of 7. Importantly, using convolution to evaluate $\mathcal{H}x$ does not require explicitly forming or storing the matrix X. The number of multiplications needed to apply the analysis filter bank using patch-based and Fourier approaches is plotted in Fig. 2.4

We also use convolution to accelerate calculation of (2.12). The first term, XX^T, is just the circular correlation of x evaluated at the first K shifts in each direction and can be calculated using FFTs at a cost of $\mathcal{O}(N^2 \log(N))$. In contrast, evaluating this gradient using a dense matrix multiply involving the image patch matrix scales as $\mathcal{O}(K^4 N^2)$. Thus the filter bank interpretation yields a savings of $K^4/\log(N)$. For typical sizes of $N = 512$ and $K = 8$, this is a 450× reduction in order. However, this term remains constant throughout the iterations and must be computed only once.

Computing the product XZ^T is more complicated. The matrix Z, while sparse, has no fixed sparsity pattern or common structure. We must compute the product with each row of Z independently using convolution. This requires a forward FFT of length N for x and for each of the N_c rows of Z, the necessary elementwise products, and finally the inverse FFT of these products. All told, this operation will scale like $\mathcal{O}(N_c N^2 \log(N))$. In contrast, directly using dense matrix multiplication scales as $\mathcal{O}(N_c K^2 N^2)$. Unlike XX^T, this term must be calculated each time z is updated.

The dominant computation in evaluating $J_1(W)$ is that of $\bar{F}W^T$. This requires N_c separate 2D-DFTs of size $N_F \times N_F$, at a cost of $\mathcal{O}(N_c N_F^2 \log N_F)$.

Similarly, calculating the gradient of $J_1(W)$ is dominated by the cost of (2.13). We first compute

Table 2.1: Computational cost for function and gradient computation.

Penalty	Evaluation	Gradient
$f(W,Z,x)$	$\mathcal{O}(N_c N^2 \log(N))$	$\mathcal{O}(N_c K^4 + N_c N^2 \log(N))$
$J_1(W)$	$\mathcal{O}\left(N_c N_F^2 \log(N_F) + N_F^2 K^2 + N_c K^2\right)$	$\mathcal{O}(K^4 N_c + (N_c + K^2) N_F^2 \log(N_F)$
$J_2(W)$	$\mathcal{O}\left(N_c^2 K^2\right)$	$\mathcal{O}(N_c K^4)$

$\bar{F} W^T$. Then, we require the multiplication of an $N_F^2 \times N_F^2$ diagonal matrix by the $N_F^2 \times K^2$ matrix $\bar{F}$ at a cost of $\mathcal{O}(N_F^2 K^2)$. Next, we take K^2 separate $N_F \times N_F$ 2D inverse FFTs, followed by the product of $N_c \times K^2$ and $K^2 \times K^2$ matrices. Together, $\nabla_W J_1(W)$ scales as $\mathcal{O}(K^4 N_c + (N_c + K^2) N_F^2 \log(N_F))$. The cost of evaluating $J_2(W)$ and $\nabla J_2(W)$ is $\mathcal{O}(N_c^2 K^2)$ and $\mathcal{O}(N_c K^4)$, respectively. These costs are summarized in Table 2.1.

For many choices of ψ, the sparse coding step is cheap. For instance, when ψ is the ℓ_0 norm, we need only to pass over each element of $\mathcal{H}x$ and set to zero all entries that are less than the given threshold. This operation will cost $\mathcal{O}(N_c N^2)$.

The necessary function and gradient evaluations consist of basic linear algebra operations, such as matrix-matrix products, and elementwise function evaluations, such as $\log(\cdot)$ or $|\cdot|^2$. As such, our algorithm is easy to implement on a graphics processing unit (GPU).

As noted, we can implement the action of the filter bank, $\mathcal{H}x$, using Fourier convolution methods, direct convolution methods, or using the patch-based multiplication WX. The best choice depends on computational platform (CPU vs. GPU), filter size, and dimensionality of training data. While Fourier methods likely win on a CPU, patch-based multiplication is well suited for GPU-based implementations. Finally, note that we can limit the amount of memory consumed by the algorithm by applying the filter bank in a channel-by-channel (or row-by-row) fashion. This is useful when learning a transform for higher-dimensional data, as the matrix WX may not fit in memory- for d-dimensional data the matrix WX is of size $N_c N^d$, while a single row of WX is of size N^d.

2.4.4 Comparison With NSOLT

The closest analogue to our proposed filter bank design algorithm is the Non-Separable Oversampled Lapped Transform (NSOLT), as described in Section 2.3.4. We briefly draw distinctions between our proposed filter bank learning algorithm and the design of multidimensional NSOLTs.

An immediate difference is that the NSOLT structure is proposed for use as a synthesis filter bank. This is not a meaningful distinction, though, as the designed NSOLT implements a tight

Algorithm 1 Filter Bank Sparsifying Transform Learning

```
INPUT: Image $x$, Initial transform $W^{(0)}$
$Z^{(0)} \leftarrow \text{prox}_{\psi\nu}\left(\mathcal{H}^{(0)}x\right)$
$k \leftarrow 0$
repeat
    $W^{(k+1)} \leftarrow \arg\min_W f(W, x, Z^{(k)}) + \mu J_1(W) + \lambda J_2(W)$
    $Z^{(k+1)} \leftarrow \text{prox}_{\psi\nu}\left(\mathcal{H}^{(k+1)}x\right)$
    $k \leftarrow k+1$
until Halting Condition
```

frame expansion, and thus the adjoint of the NSOLT is itself a paraunitary analysis filter bank with compactly supported FIR filters [12].

The true differences between our approach and the design of NSOLTs lie in the structure of the filter bank. First, our algorithm is only applicable to undecimated filter banks, while NSOLTs can incorporate downsampling. Second, our approach can learn any undecimated PR filter bank with compactly supported FIR filters, whereas the NSOLT framework can learn only a subset of paraunitary filter banks. This difference manifests itself in both the structure and optimization of the filter banks. Our filter banks are unstructured, and use special regularizers to ensure the PR property holds. In our approach, the frame bounds are indirectly controlled through the penalty parameter μ. In contrast, the NSOLT uses a particular lattice form that implicitly guarantees the learned filter bank is paraunitary.

Further, NSOLTs are designed using a combination of symmetric and anti-symmetric impulse responses to ensure the filter bank is linear phase [65–67]. This constraint limits the ability of individual NSOLT channels to capture structures which are not strictly symmetric or anti-symmetric, such as edges that are not strictly horizontal, vertical, or at an angle of 45 degrees. Our undecimated filter banks do not have this restriction.

2.5 Application to Image Denoising

A preliminary version of our filter bank transform learning framework has been applied in an “adaptive” manner for magnetic resonance imaging [86]. Here, we restrict our attention to image denoising in the “universal” paradigm: we use a pre-trained sparsifying filter bank, $\mathcal{H}$ to recover a clean image x^* from a noisy copy, $y = x^* + e$, where $e \sim \mathcal{N}(0, \sigma^2 I_N)$. We consider two algorithms for image denoising.

Algorithm 2 Iterative Denoising With Filter Bank Transform

```
INPUT: Noisy signal $y$, Learned filter bank transform $\mathcal{H}$
$k \leftarrow 0$
repeat
    $z^{(k+1)} \leftarrow \operatorname{prox}_{\psi\nu}\left(\mathcal{H}x^{(k)}\right)$
    $x^{(k+1)} \leftarrow (\mathcal{H}^*\mathcal{H} + \lambda_r I)^{-1}(\mathcal{H}^* z^{(k)} + \lambda_r y)$
    $k \leftarrow k+1$
until Halting Condition
```

2.5.1 Iterative Denoising

Our first denoising method is to solve a regularized inverse problem using a transform sparsity penalty, written as

$$\min_{x,z} \frac{\lambda_r}{2}\|y-x\|_2^2 + \frac{1}{2}\|\mathcal{H}x - z\|_2^2 + \nu\psi(z),$$

where $\lambda_r > 0$ controls the regularization strength. We solve this problem by alternating minimization: we update z for fixed x, and then update x with z fixed. This procedure is summarized as Algorithm 2. The eigenvalue decomposition of Lemma 2.1 provides an easy way to compute the necessary matrix inverse for cyclic convolution filter banks. For linear convolution filter banks, we use Lemma 2.1 to implement a circulant preconditioner [16].

Algorithm 2 has three key parameters. The regularization parameter λ_r reflects the degree of confidence in the noisy observations y and should be chosen inversely proportional to the noise variance. The sparsity of the transform sparse code is controlled by ν. The value of ν when denoising an image need not be the same as ν during the learning procedure and should be proportional to σ. The choice of both ν and λ_r depends on the final parameter: the number of iterations used during denoising. Empirically, we've found that using $\operatorname{ceil}\{\sigma \cdot 255/10\}$ iterations works well.

2.5.2 Denoising by Transform-Domain Thresholding

We also consider a simpler algorithm, inspired by the transform domain denoising techniques of old. We can form a denoised estimate by passing y through the system in Fig. 2.3; that is, computing

$$\mathcal{H}^\dagger \operatorname{prox}_{\psi\nu}\left(\mathcal{H}y\right). \tag{2.14}$$

This approach simplifies denoising by eliminating two parameters from Algorithm 2: the number of iterations and λ_r.

Denoising in this manner is sensible because of the properties we have imposed on $\mathcal{H}$. Noise in

Figure 2.5: Training images.

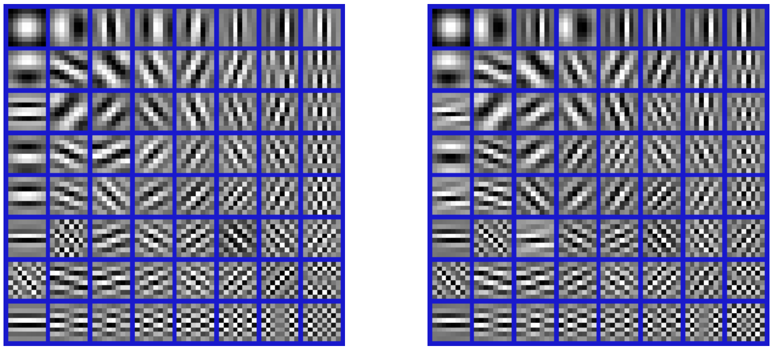
the signal will not be sparse in the transform domain and thus will be reduced by the nonlinearity. In contrast, the image will be sparse in the transform domain, and significant components will pass through the nonlinearity with little change. The left-inverse is guaranteed to exist, and as $\mathcal{H}$ must be well-conditioned, any noise remaining after the nonlinearity will not be strongly amplified by $\mathcal{H}^\dagger$. Finally, if $\mathcal{H}$ has low coherence, the transformed noise $\mathcal{H}e$ will not be correlated across channels, suggesting that a channelwise nonlinearity is sufficient. A multi-channel nonlinearity may be beneficial if the transform is coherent.

Unless $\mathcal{H}$ implements a tight frame expansion, the minimum norm synthesis filters comprising $\mathcal{H}^\dagger$ will not be compactly supported; indeed, if $\mathcal{H}$ is a linear convolution filter bank, then the minimum-norm synthesis filter bank will have infinite duration impulse response filters [12]. Fortunately, if $\mathcal{H}$ is well-conditioned, then the minimum-norm synthesis filters will have an exponentially decaying impulse response and can thus be well approximated by FIR filters [38]. Alternatively, one can search for a (non-minimum-norm) left inverse of $\mathcal{H}$ that consists of FIR filters [37].

2.6 Experiments

We implemented GPU versions of our algorithms using NumPy 1.11.3 and SciPy 0.18.1. Our code interfaces with Python through `PyCUDA` [87] and `scikits-cuda` [88], and we conducted experiments on an NVidia Maxwell Titan X GPU.

We conducted training experiments using the five training images in Fig. 2.5. Each image, in testing and training, was normalized to have unit ℓ_2 norm. Unlike many patch-based methods, we do not subtract the DC (mean) value of the image prior to training. Unless otherwise specified, our transforms were learned using 1000 iterations of Algorithm 1 with parameters $\mu = 3.0$, $\lambda = 7 \times 10^{-4}$, and $\nu = 5.5 \times 10^{-3}$. Sparsity was promoted using an ℓ_0 penalty, for which the prox operator corresponds to hard thresholding. For each filter bank, we compute the coherence between each pair of squared magnitude filter responses and report the largest value; that is, $\max_{1 \le i < j < N_c} \Gamma_{i,j}(\left|\bar{F}W^T\right|^2)$. The initial transform $\mathcal{H}^{(0)}$ must be feasible, *i.e.* left-invertible. Random Gaussian and DCT initializations work well in practice. We learned a 64-channel



(a) (b)

Figure 2.6: Comparing initializations for 64-channel filter bank with 8 × 8 filters. (a) DCT initialization; maximum coherence: 0.9 (b) Random initialization; maximum coherence: 0.85. The filters in (b) have been ordered to match those in (a).

filter bank with 8 × 8 filters using these initializations. The learned filters are shown in Fig. 2.6. The evolution of the objective function and sparsification error are shown in Fig. 2.7. The learned filters appear similar, reach nearly the same objective value, and perform equally well in sparsifying the data set.

Additional examples of learned filters and their magnitude frequency responses are shown in Fig. 2.8. We show a subset of channels from a filter bank consisting of 16 × 16 filters and 128 channels. This transform is 2× *under*-complete if viewed as a patch-based transform. The ability to choose longer filters without increasing the number of channels is a key advantage of our framework over patch-based transform learning.

2.6.1 Image Denoising

We investigate the denoising performance of the filter bank sparsifying transforms as a function of number of channels, N_c, and filter size, K, using our two algorithms. We refer to Filter Bank Sparsifying Transform (FBST) with N_c channels and $K \times K$ filters as FBST-N_c-K.

We evaluate our filter bank learning formulation using 64, 128, and 256 channels with 8 × 8 and 16 × 16 filters. During the denoising stage, we set $\nu = 10^{-4} \times 0.1\sigma$ and λ_r was adjusted for the particular noise level.

We also evaluate image denoising using filters learned with the square, patch-based transform learning algorithm [41]. We used 8×8 and 16×16 image patches to learn a patch-based transform W. We used the rows of W to generate an undecimated filter bank and used this filter bank to denoise using Algorithm 2 and (2.14). The filter bank implements cyclic convolution and image

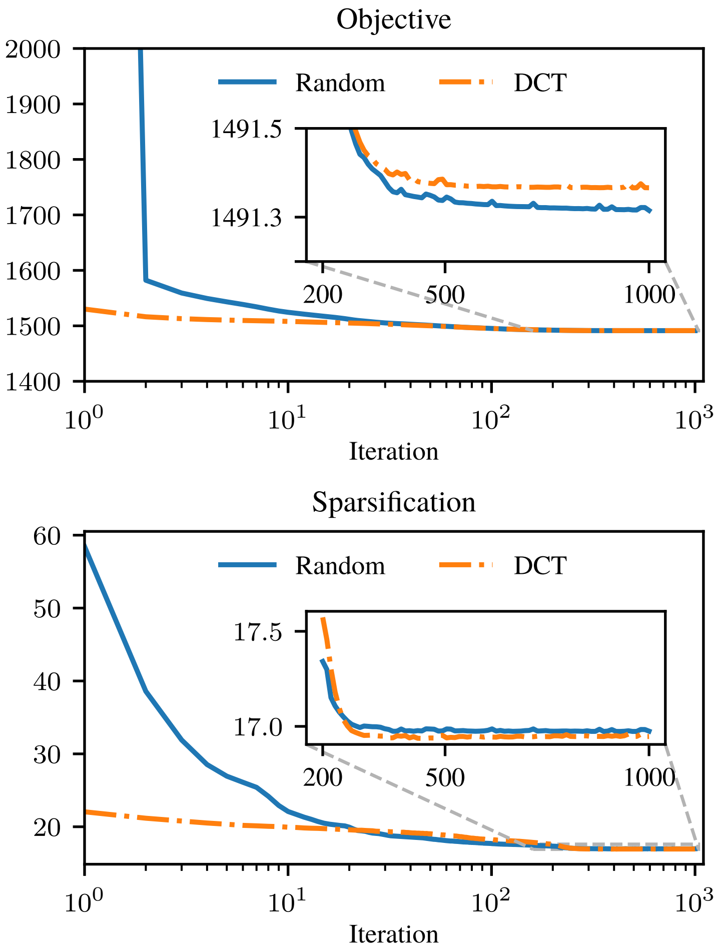


Figure 2.7: Plots of objective function (2.10) and sparsification error $\frac{1}{2}\|\mathcal{H}x - z\|_2^2$ while training the filter banks shown in Fig. 2.6.

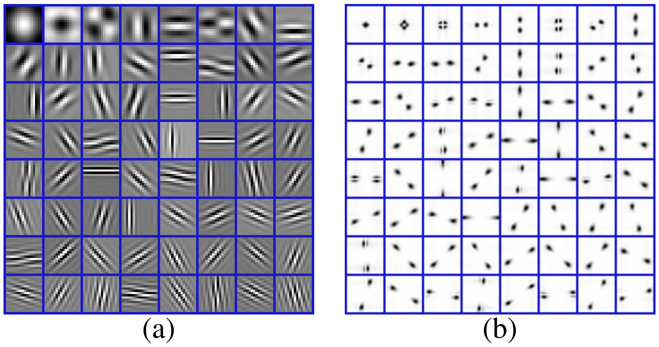


Figure 2.8: Examples of learned 16 × 16 filters. (a) Filter impulse responses; (b) Magnitude frequency responses. The zero frequency (DC) is located at the center of each small box. Maximum coherence of learned filter bank: 0.88.

patches are extracted using periodic boundary conditions. Following the convention used to denote filter bank sparsifying transforms, we refer to a Patch-Based Sparsifying Transform (PBST) with $K \times K$ patches as PBST-K^2-K. We keep the number of channels $N_c = K^2$ explicit to facilitate comparison with filter bank sparsifying transforms. Observe that for a given K, FBST-K^2-K and PBST-K^2-K differ only in the regularizer and learning algorithm used; in particular, both transforms have the same number of design parameters. All sparsifying transforms were trained under the "universal" paradigm using all possible (maximally overlapping) patches extracted from the set of images shown in Fig. 2.5.

We also include comparisons with 2D NSOLTs trained using the `SaivDr`[2] MATLAB package. The NSOLTs were trained using the images shown in Fig. 2.5. We investigate denoising performance as a function of polyphase order, downsampling ratio, and number of channels. We refer to an NSOLT with downsampling matrix $2I$, 48 channels, and polyphase order 4 as NSOLT-2-48-4. We consider only NSOLTs with an identical number of symmetric and antisymmetric channels, thus NSOLT-2-48-4 has 24 symmetric and 24 anti-symmetric channels.

Image denoising using NSOLTs is accomplished by solving $\arg\min_x \|y-x\|^2/2+\lambda\|\mathcal{H}x\|_1$, where $\mathcal{H}$ denotes the NSOLT operator and λ is tuned for best denoising performance on an image-by-image basis. The optimization problem itself was solved using FISTA [89]. We also attempted denoising by using Algorithm 2 with an ℓ_0 penalty, but found that ℓ_1 + FISTA gave the best results. We used two tree levels for each NSOLT.

While our main interest is comparing the denoising performance of FBST versus PBST, we also compare against several competing image denoising methods. These are divided into two camps. The first group includes two methods based on non-local self-similarity: BM3D [90] and WNNM [91].

The second group includes a handful of MRF learning-based methods. These can be interpereted as either patch-based analysis or convolutional analysis models and are thus local in nature. We include EPLL-GMM [92], the Field of Experts (FoE) [56, 93], and the Cascade of Shrinkage Fields (CSF) [78]. While these methods have a convolutional structure– indeed, FoE and CSF are closely related to our nonlinear analysis-synthesis filter bank– they must be trained in a supervised in nature and do not impose any perfect reconstruction property on the resulting filters. We use the default parameters in the FoE and CSF packages. For FoE, we use 3 × 3 filters. For CSF we use 5 × 5 and 7 × 7 filters; in both cases, we use five stages with 25 channels. We trained the CSF to operate at our noise levels.

Finally, we include the recent STROLLR denoising algorithm, which uses both square patch-based transform learning and non-local self-similarity [94, 95]. STROLLR is an unsupervised

[2] Available: `https://github.com/msiplab/SaivDr`

method and the transform learning step uses the "adaptive" paradigm. We used 8 × 8 patches, resulting in a 64 × 64 sparsifying transform.

Our metric of interest is the peak signal-to-noise ratio (PSNR) between the reconstructed image x and the ground truth image x^*, defined in decibels as PSNR $= 20\log_{10}(N^2/\|x - x^*\|_2)$. We evaluate the denoising performance of our algorithm on the grayscale `barbara`, `man`, `peppers`, `baboon` and `boat` images.

Table 2.2 collects the reconstruction PSNR for each test image, in addition to the mean PSNR for the entire test set. The best value is written in bold with gray shading; the second-best value is shaded gray with no bold. Here, FBST-64-8 indicates a 64-channel filter bank with 8 × 8 filters where we denoise using transform domain thresholding (2.14), while FBST-128-16-I indicates the use of a 128-channel filter bank with 16 × 16 filters and denoising using the iterative Algorithm 2.

When averaged over the entire test set, FBST-64-8 outperforms PBST-64-8 by between 0.2 – 0.3 dB. As the only difference in these two transforms is the regularizer used during learning, we attribute this improvement to the change from a patch-based to an image-based point of view.

Using 1000 iterations to learn a 196-channel filter bank with 16 × 16 filters with Algorithm 1 took roughly five minutes on our GPU. In contrast, using the same GPU to learn a square 256 × 256 patch-based transform over the same data took less than one minute. This illustrates the efficiency of the closed-form transform update step in the patch-based case [43]. Our slower learning algorithm is offset by the ability to choose $N_c < K^2$, and this leads to faster application of the learned transform. Comparing FBST-128-16-I and PBST-256-16-I, the image-based transform outperforms the patch-based transform by up to 0.3 dB despite containing half as many channels.

Table 2.2 shows that, on average, FBST performs slightly better than the MRF-based methods (EPLL/CSF/FoE), but worse than non-local methods, especially for $\sigma = 30$. Note that STROLLR is competitive with the other non-local methods. Combining the flexibility of our proposed filter bank sparsifying transforms with STROLLR is left for future work. NSOLT gives the lowest denoising performance. We conjecture that the linear phase constraints severely limit the representation power of the NSOLT. Further, it is difficult to train a large NSOLT: training NSOLT-2-24-2 required several days on our workstation.

The performance of shorter or longer filters is dependent on the image. For most images, the 8 × 8 filters performed as well or better than the longer filters, but we see significant improvement when using long filters on `barbara`. The MRF-based methods perform poorly on this image, with the FBST-I methods gaining well over a full dB of PSNR improvement. In contrast, the MRF methods outperform FBST on `man`.

Increasing the number of channels beyond the filter size K^2 provides marginal improvement.

Table 2.2: Reconstruction PSNR for test images averaged over 10 noise realizations. The column **mean** reports the mean PSNR over the set of test images. The highest PSNR in each column shaded gray and in bold; the second highest result is shaded gray. FBST-128-16 indicates a filter bank sparsifying transform with 128 channels and 16 × 16 filters and denoised according to (2.14). The -I suffix indicates denoising with the iterative Algorithm 2. CSF$_{7\times7}$ indicates a cascaded shrinkage field with 7 × 7 filters. FoE$_{3\times3}$ denotes the Field of Experts using 3 × 3 filters. NSOLT-2-48-2 indicates an NSOLT with downsampling by 2, 48 channels, and polyphase order 2.

σ	10	20	30	10	20	30	10	20	30	10	20	30	10	20	30
Input PSNR	28.1	22.1	18.6	28.1	22.1	18.6	28.1	22.1	18.6	28.1	22.1	18.6	28.1	22.1	18.6
Method	**mean**			**baboon**			**barbara**			**man**			**peppers**		
WNNM	**33.8**	**30.6**	**28.8**	**30.7**	**26.7**	24.6	**35.4**	**32.2**	**30.3**	**34.2**	**30.7**	**28.9**	**34.8**	**32.6**	**31.1**
BM3D	33.6	30.4	28.6	30.5	26.5	24.4	35.0	31.7	29.8	34.0	30.6	28.8	**34.8**	32.5	31.0
STROLLR	33.6	30.4	28.6	30.5	26.6	**24.7**	35.1	31.9	29.9	33.8	30.5	28.7	**34.8**	32.4	30.9
EPLL	33.3	29.9	28.1	30.5	26.6	24.6	33.6	29.7	27.5	33.9	30.6	28.8	34.6	32.3	30.6
CSF$_{5\times5}$	33.1	29.7	27.7	30.3	26.3	24.1	33.4	29.3	26.8	33.8	30.4	28.6	34.6	32.0	30.3
CSF$_{7\times7}$	33.1	29.7	27.7	30.4	26.3	24.2	33.4	29.3	26.8	33.7	30.4	28.6	34.6	32.1	30.3
FOE$_{3\times3}$	32.8	29.1	27.1	30.1	25.9	23.7	32.6	28.1	25.5	33.5	30.0	28.2	34.3	31.5	29.8
PBST-64-8	33.1	29.5	27.5	30.3	26.0	23.8	34.0	30.0	27.8	33.4	29.8	28.0	34.4	31.7	29.9
PBST-256-16	33.3	29.7	27.7	30.4	26.1	23.9	34.2	30.3	28.0	33.6	30.0	28.0	34.6	31.9	30.1
PBST-64-8-I	33.2	29.8	27.8	30.3	26.1	24.1	34.1	30.3	28.0	33.6	30.1	28.3	34.6	32.0	30.4
PBST-256-16-I	33.4	29.9	27.9	30.4	26.4	24.2	34.2	30.5	28.2	33.8	30.2	28.4	34.7	32.2	30.5
FBST-64-8	33.3	29.8	27.8	30.3	26.1	23.9	34.4	30.5	28.2	33.7	30.1	28.2	34.5	31.9	30.2
FBST-128-8	33.3	29.8	27.8	30.4	26.1	24.0	34.3	30.5	28.3	33.7	30.1	28.3	34.6	31.9	30.2
FBST-196-8	33.3	29.8	27.8	30.4	26.2	24.0	34.3	30.5	28.2	33.7	30.2	28.2	34.5	31.9	30.1
FBST-64-16	33.3	29.8	27.8	30.2	26.0	23.9	34.5	30.8	28.6	33.5	30.0	28.1	34.4	31.9	30.1
FBST-128-16	33.3	29.9	28.0	30.3	26.1	24.0	34.6	31.0	28.8	33.6	30.1	28.2	34.6	32.0	30.4
FBST-196-16	33.4	30.0	28.0	30.3	26.1	24.0	34.7	31.1	29.0	33.6	30.1	28.2	34.6	32.0	30.4
FBST-64-8-I	33.4	30.0	28.1	30.4	26.3	24.2	34.4	30.7	28.5	33.8	30.3	28.5	34.7	32.2	30.6
FBST-128-8-I	33.4	30.0	28.1	30.4	26.4	24.3	34.3	30.7	28.5	33.8	30.3	28.4	34.7	32.2	30.6
FBST-196-8-I	33.4	29.9	28.0	30.5	26.4	24.2	34.3	30.6	28.3	33.8	30.3	28.4	34.6	32.1	30.5
FBST-64-16-I	33.4	30.1	28.1	30.4	26.3	24.2	34.6	31.1	29.0	33.6	30.3	28.5	34.7	32.2	30.6
FBST-128-16-I	33.4	30.1	28.3	30.4	26.4	24.3	34.7	31.2	29.2	33.7	30.2	28.5	34.7	32.3	30.7
FBST-196-16-I	33.5	30.2	28.3	30.4	26.3	24.4	34.8	31.4	29.3	33.7	30.3	28.4	34.7	32.3	30.8
NSOLT-4-12-2	30.3	26.0	23.8	29.1	24.3	21.9	30.1	25.7	23.4	30.5	26.6	24.5	31.2	27.1	24.9
NSOLT-2-12-2	30.9	26.8	24.7	29.3	24.6	22.4	30.8	26.4	24.2	31.3	27.3	25.3	32.0	28.3	26.2
NSOLT-4-12-4	31.0	26.8	24.7	29.3	24.6	22.4	30.9	26.5	24.3	31.3	27.3	25.3	32.1	28.4	26.3
NSOLT-4-24-2	30.8	26.6	24.5	29.2	24.5	22.3	30.9	26.7	24.4	31.0	27.0	25.0	31.8	27.9	25.9
NSOLT-4-24-4	31.1	27.0	24.9	29.3	24.7	22.5	31.0	26.7	24.4	31.4	27.6	25.6	32.2	28.6	26.6
NSOLT-2-12-4	29.9	25.6	23.4	28.9	24.1	21.7	29.8	25.4	23.2	30.1	26.0	24.1	30.8	26.6	24.5

For low noise, denoising by transform-domain thresholding and iterative denoising using Algorithm 2 perform equally well. As the noise level increases, the iterative denoising algorithm outperforms the simpler thresholding scheme.

2.6.2 Learning on a Subset of Patches

One advantage of patch-based formulation is that the model can be trained using a large set images by randomly selecting a few patches from each image. We can use the same approach when learning a sparsifying filter bank: the data matrix X in (2.7) is formed by extracting and vectorizing patches from many images. We can no longer view WX as a convolution.

We learned a transform using $200,000$ randomly extracted patches from the training images in Fig. 2.5. The learned transform performed nearly identically to a transform learned using all patches from the training images.

2.6.3 Image Adaptivity

To test the influence of the training set, we learned a filter bank using 256^2 patches of size 8×8 chosen at random from the 200 training images in the BSDS300 training set [96]. The learned filter bank consists of Gabor-like filters, much like filter banks learned from the images in Fig. 2.5. Gabor-like filters are naturally promoted by the regularizers J_1 and J_2: their narrow support in the frequency domain leads to low coherence, and their magnitude responses can tile frequency space leading to a well-conditioned transform. As expected, all but one filter has zero-mean.

We wondered if we have regularized our learning problem so strongly that the data no longer plays a role. Fortunately, this is not so: Fig. 2.9 illustrates a 64 channel filter bank of 16×16 filters learned from a highly symmetric and geometric image. The learned filters include oriented edge detectors, as in the natural image case, but also filters with a unique structure that sparsify the central region of the image. Note that most of our learned filters in Figs. 2.8 and 2.9 are not strictly symmetric or antisymmetric, and thus cannot be captured by individual NSOLT channels.

2.7 Remarks

Adaptive analysis/transform sparsity based image denoising algorithms can be coarsely divided into two camps: supervised and unsupervised. In both cases, one learns a signal model by minimizing an objective function.

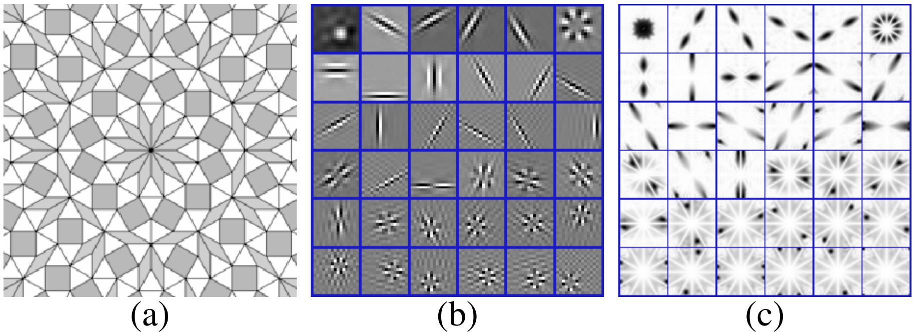


Figure 2.9: Adaptivity of filters. (a) Training image; (b) Learned filter impulse responses. (c) Magnitude frequency responses. Zero frequency (DC) is located at the center of each small box. Maximum coherence of learned filter bank: 0.75.

In the supervised case, this minimization occurs over a set of training data. In a denoising application one typically corrupts a clean input with noise, passes it through the denoiser, and uses the difference between the clean and denoised signal to adapt various components of the denoising algorithm: the analysis operator, thresholding parameters, mixture weights, the number of iterations, and so on. It is not necessary to regularize the learning procedure to preclude degenerate solutions, such as a transform of all zeros; such a transform would not perform well at the denoising task, and thus would not be learned by the algorithm [56, 57, 68].

In the unsupervised case, the objective function has two components. The first is a surrogate for the true, but unavailable, metric of interest. In this chapter, we use the combination of sparsity and sparsification error to act as a surrogate for reconstruction PSNR. The second part of the objective is a regularizer that prevents degenerate solutions, as discussed in Sections 2.2.2 and 2.2.3. Even in the "universal" case, our learning is essentially unsupervised, as the learning process is not guided by the denoising PSNR.

The TNRD algorithm [58] is a supervised approach that resembles iterative denoising using Algorithm 2, but where the filter coefficients, nonlinearities, and regularization parameter are allowed to vary as a function of iteration. However, the TNRD approach has no requirements that the filters form a well-conditioned frame or have low coherence; "poor" filters are simply discouraged by the supervised learning process. Denoising with the TNRD algorithm outperforms the learned filter bank methods presented here.

One may ask if it is necessary that the learned transform be a frame. Indeed, the matrix to be inverted when denoising using Algorithm 2 is full-rank even if the filter bank itself is not perfect reconstruction. The proposed regularizer, while less restrictive than previous transform learning regularizers, may still overly constrain the set of learnable sparsifying transforms. However, our highly regularized approach has a benefit of its own. Whereas the TNRD algorithm is trained on hundreds of images, and can take upwards of days to train, our algorithm can be applied to a single image and requires only a few minutes. The tradeoff offered by the TNRD algorithm is acceptable for image denoising tasks, as massive sets of natural images are publicly available for use in machine learning applications. However, such data sets may not be available for new imaging modalities, in which case a tradeoff closer to that offered by our filter bank learning algorithm may be preferred. Finding a balance between our highly regularized and unsupervised approach and competing supervised learning methods is the subject of ongoing work.

2.8 Conclusions

We have developed an efficient method to learn sparsifying transforms that are structured as undecimated, multidimensional perfect reconstruction filter banks. Unlike previous transform learning algorithms, our approach can learn a transform with fewer rows than columns. We anticipate this flexibility will be important when learning a transform for high-dimensional data. Numerical results show our filter bank sparsifying transforms outperform existing patch-based methods in image denoising. Future work might fully embrace the filter bank perspective and learn filter bank transforms with various length filters and/or non-square impulse responses.

Chapter 3

Interferometric Synthetic Aperture Microscopy

3.1 Introduction

In this chapter, we review Interferometric Synthetic Aperture Microscopy (ISAM). This material is the starting point for Chapters 4 and 5.

ISAM is a noninvasive, scattering-based imaging modality that reconstructs the 3D spatial distribution of a target from interferometric measurements using a two-dimensional (planar) scanning geometry.

ISAM is closely linked to Optical Coherence Tomography (OCT). In OCT, the target is illuminated with focused, broadband light. This light interacts with the target and is scattered back into the instrument. The instrument is scanned along a two-dimensional planar trajectory, and the experiment is repeated at each point. OCT relies on a pencil-beam approximation to the illumination beam, resulting in a loss of transverse resolution away from the focal plane. The pencil-beam approximation results in an tradeoff between depth-of-field and transverse resolution, as highly focused beams do not satisfy the pencil-beam approximation.

The combination of OCT and highly focused beams is known as Optical Coherence Microscopy (OCM). OCM requires volumetric scanning (that is, 3D spatial scanning) in addition to broadband measurements to obtain appreciable depth of field.

The ISAM forward model removes the pencil-beam approximation and models the diffraction of the illuminating field away from the focal plane. By solving the inverse scattering problem, ISAM obtains depth-invariant resolution without the need for volumetric scanning.

3.2 Notation

Here, and in the rest of this thesis, we write the set of integers $\{1, 2, \ldots, N\}$ as $[N]$ and the imaginary unit as i. Linear mappings between Hilbert spaces are written in calligraphic font, *e.g.* $\mathcal{A}$. The adjoint of $\mathcal{A}$ is written $\mathcal{A}^*$. Finite-dimensional vectors are denoted by lower-case bold letters, *e.g.*

$\mathbf{x} \in \mathbb{C}^N$. Finite-dimensional matrices and tensors are written using upper-case bold letters. We adopt Matlab-style indexing notation: given a matrix $\mathbf{A} \in \mathbb{C}^{N \times M}$, its i-th row is $\mathbf{A}[i, :]$, the j-th column is $\mathbf{A}[:, j]$, and i, j-ith element is $\mathbf{A}[i, j]$. We denote the vector $\text{vec}(\mathbf{A}) \in \mathbb{C}^{NM}$ is formed by stacking the columns of $\mathbf{A}$ into a single vector (*i.e.*, row-major ordering). The range, null space, and rank of a matrix $\mathbf{A}$ are written $\text{range}\{\mathbf{A}\}, \text{null}\{\mathbf{A}\}$, and $\text{rank}\{\mathbf{A}\}$. Given $\mathbf{x} \in \mathbb{C}^N$, the diagonal matrix $\text{diag}(\mathbf{x}) \in \mathbb{C}^{N \times N}$ has the entries of $\mathbf{x}$ along its main diagonal. Similarly, given a set of $N \times M$ matrices $\mathbf{A}_1, \ldots, \mathbf{A}_L$, the matrix $\text{blkdiag}(\mathbf{A}_1, \ldots, \mathbf{A}_L) \in \mathbb{C}^{LN \times LM}$ is block-diagonal with the collection of $\mathbf{A}_i$ along its block diagonal.

The transpose (*resp.* Hermitian transpose) of a matrix is written $\mathbf{A}^\mathsf{T}$ (*resp.* $\mathbf{A}^\mathsf{H}$). The ℓ_p norm of $\mathbf{x} \in \mathbb{C}^N$ is $\|\mathbf{x}\|_p = \left(\sum_{j=1}^N \left|\mathbf{x}[j]\right|^p\right)^{1/p}$. For vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$ we use the shorthand $|r| = \|\mathbf{r}\|_2$. The $N \times N$ identity matrix is $\mathbf{I}_N$, and the vector $[1, 1, \ldots 1]^\mathsf{T} \in \mathbb{R}^N$ is written $\mathbb{1}_N$. The tensor (or Kronecker) product between matrices $\mathbf{A}$ and $\mathbf{B}$ is $\mathbf{A} \otimes \mathbf{B}$.

Let X be a (nonempty) set. $L^2(X)$ denotes the Hilbert space of functions with domain X that are square integrable functions with respect to the Lebesgue measure. The inner product on $L^2(X)$ is $\langle f, g\rangle_{L^2(X)} = \int_X f g^* \mathrm{d}\mu$ where μ is the Lebesgue measure. The norm on $L^2(X)$ is written $\|f\|_{L^2(X)}$. The space of functions with p continuous derivatives and domain X is written $C^p(X)$.

3.3 ISAM Forward Model

Throughout this dissertation, we restrict our attention to the scalar field model. We model the sample through its complex refractive index, $n(\mathbf{r}, k_0) = n_b + \delta n(\mathbf{r}, k_0)$ where n_b is the refractive index of the background medium and δn is the perturbation due to the sample; for simplicity, we take $n_b = 1$. Here, $\mathbf{r} = (x, y, z) = (\mathbf{r}_\parallel, z)$, where $\mathbf{r}_\parallel$ are the transverse dimensions and z indicates the axial dimension. We assume that δn is (spatially) supported in the bounded region $\Gamma \subset \mathbb{R}^3$. The free-space wavenumber k_0 is related to temporal angular frequency ω by $k_0 = \omega / c$, where c is the speed of light in free space. The real part of the complex refractive index is the ratio between c and the phase velocity in the medium, while the imaginary part indicates attenuation due to propagation through the target.

Under the first Born approximation, which applies for semi-transparent or weakly scattering objects, the obtained measurements are linear in the complex susceptibility $\eta \triangleq n^2 - 1$; we will work with the susceptibility rather than the refractive index. Note that η is also supported on Γ.

In the context of spectroscopy, the “spectrum” of a sample usually refers either to its complex refractive index or only to the imaginary part of the refractive index. Consider a homogeneous medium with refractive index $n(k_0) = n_r(k_0) + \mathrm{i}\kappa(k_0)$. The real part, $n_r(k_0)$, has mean value

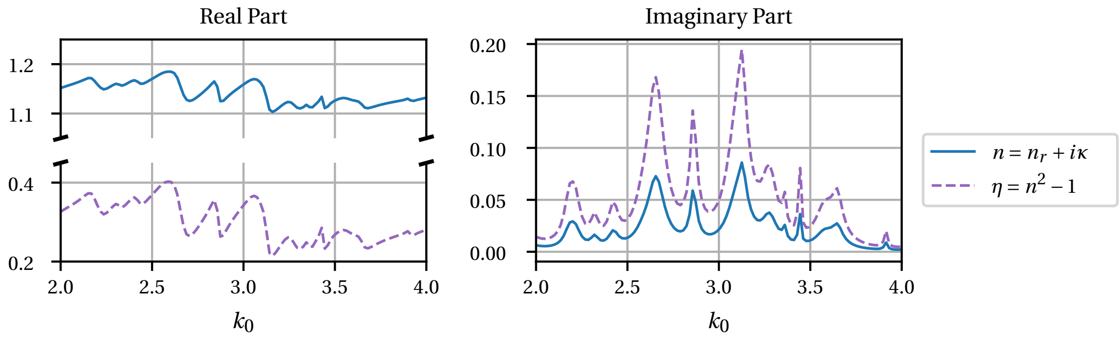


Figure 3.1: Comparing the complex refractive index and complex susceptibility.

greater than one and the imaginary part, $\kappa(k_0)$, is non-negative. Relating $\eta(k_0)$ to $n(k_0)$, we have

$$\eta(k_0) = n(k_0)^2 - 1 = n_r(k_0)^2 - \kappa(k_0)^2 - 1 + 2\mathrm{i} n_r(k_0)\kappa(k_0).$$

Unlike the refractive index, the mean value of the real part of $\eta(k_0)$ may be less than one and can be negative. The imaginary part of $\eta(k_0)$ remains non-negative. A comparison of $n(k_0)$ and $\eta(k_0)$ is shown in Fig. 3.1.

The ISAM imaging geometry is shown in Fig. 3.2. We consider a confocal point scanning system where the illuminating aperture serves as the detection aperture.

The aperture is located in the plane $z = 0$. With the aperture positioned at $(\mathbf{r}_{\|}^{(o)}, 0)$, the sample is illuminated by a broadband Gaussian beam focused to a point $\mathbf{r}^{(o)} = (\mathbf{r}_{\|}^{(o)}, z_F)$ within the sample. The illuminating field interacts with the sample, and a portion of the light is scattered backwards and is collected through the aperture. The aperture is raster scanned (either optically or mechanically) along the transverse coordinates $\mathbf{r}_{\|}^{(o)}$. At each point the scattered field is measured interferometrically, from which we use standard techniques to recover the complex (phase-resolved) measurements. In this thesis, we ignore the interferometric aspects of data acquisition and work directly with the phase-resolved measurements.

The illuminating field, $u_i(\mathbf{r}, k_0)$, takes the form of a Gaussian beam centered at $\mathbf{r}_{\|}^{(o)}$ and focused to a depth z_F within the sample. In the focal plane, the transverse Fourier transform of u_i is a Gaussian function:

$$\begin{aligned}\hat{g}(\mathbf{k}_{\|}, k_0) &= \frac{\rho(k_0)}{2\pi} \int u_i(\mathbf{r}_{\|}, z_F, k_0) e^{-\mathrm{i}\mathbf{k}_{\|}\cdot\mathbf{r}_{\|}}\, \mathrm{d}^2 r_{||} \\ &= \frac{\rho(k_0)}{k_0 \mathrm{NA}} \exp\left\{-\frac{\left|\mathbf{k}_{\|}\right|^2}{(k_0\mathrm{NA})^2}\right\}, \end{aligned} \tag{3.1}$$

where $\rho(k_0)$ is the power spectrum of the broadband illuminating source. We assume that

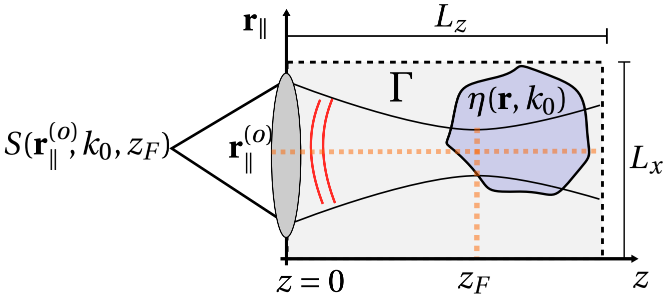


Figure 3.2: Geometry and notation for the scattering problem under consideration. The illuminating aperture is located at $(\mathbf{r}_\parallel^{(o)}, 0)$. The incident field emerges from the aperture and is focused to the plane $z = z_F$. The incident beam interacts with the sample, η, and the backscattered light (red) is collected through the aperture to produce the measurement $S(\mathbf{r}_\parallel^{(o)}, k_0, z_F)$.

$\rho(k_0)$ is supported on the interval $[k_a, k_b]$. The scalar NA denotes the numerical aperture of the illumination system, defined to be the sine of the angle at which the Gaussian beam falls to e^{-1} of its maximum value [97].

Under the first Born approximation, the measured data is a linear function of η; we have

$$S(\mathbf{r}_\parallel^{(o)}, z_F, k_0) = \iint A(\mathbf{r}_\parallel^{(o)} - \mathbf{r}_\parallel, z - z_F, k_0)\eta(\mathbf{r}_\parallel, z, k_0)\, \mathrm{d}z\, \mathrm{d}^2 r_\parallel, \tag{3.2}$$

or, after taking a Fourier transform along the scanning dimension $\mathbf{r}_\parallel^{(o)}$,

$$\hat{S}(\mathbf{k}_\parallel, z_F, k_0) = \frac{1}{2\pi}\int S(\mathbf{r}_\parallel^{(o)}, z_F, k_0) e^{-\mathrm{i}\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel^{(o)}}\, \mathrm{d}^2 r_\parallel^{(o)} = \int \hat{A}(\mathbf{k}_\parallel, z - z_F, k_0)\hat{\eta}(\mathbf{k}_\parallel, z, k_0)\mathrm{d}z. \tag{3.3}$$

We call the function $\hat{A}$ the *ISAM kernel.* Explicitly,

$$\hat{A}(\mathbf{k}_\parallel, z, k_0) \triangleq \left|\rho(k_0)\right|^2 \int_{\Omega(\mathbf{k}_\parallel, k_0)} \frac{\hat{g}(\mathbf{k}'_\parallel, k_0)\hat{g}(\mathbf{k}_\parallel - \mathbf{k}'_\parallel, k_0)}{k_z(\mathbf{k}'_\parallel, k_0)} e^{\mathrm{i}\left(k_z\left(\mathbf{k}'_\parallel, k_0\right) + k_z\left(\mathbf{k}_\parallel - \mathbf{k}'_\parallel, k_0\right)\right)(z - z_F)}\, \mathrm{d}^2 k'_\parallel, \tag{3.4}$$

where

$$k_z(\mathbf{k}_\parallel, k_0) \triangleq \sqrt{k_0^2 - \left|\mathbf{k}_\parallel\right|^2}$$

and the set $\Omega(\mathbf{k}_\parallel, k_0) \triangleq \left\{\mathbf{k}'_\parallel \in \mathbb{R}^2 : \left|\mathbf{k}_\parallel - \mathbf{k}'_\parallel\right| \leq k_0,\ \left|\mathbf{k}'_\parallel\right| \leq k_0\right\} \subset \mathbb{R}^3$ restricts the integral to propagating modes.

3.4 Approximations to the ISAM Kernel

Next, we discuss the ISAM inverse problem—recovery of the object η from measurements of the form (3.3).

First, note that $\rho(k_0)$ in (3.4) ensures that $\hat{A}(\mathbf{k}_\parallel, z, k_0)$ vanishes for any $k_0 \notin [k_a, k_b]$. Further, $\Omega(\mathbf{k}_\parallel, k_0)$ is empty for $\left|\mathbf{k}_\parallel\right| > 2k_0$ and so $\hat{A}(\mathbf{k}_\parallel, z, k_0)$ vanishes for all $\left|\mathbf{k}_\parallel\right| > 2k_b$. This is a consequence of the diffraction limit [98].

Previous work solved the ISAM inverse problem using a perturbative approach [99–102]. The ISAM kernel is replaced by a simpler approximation and the simplified problem is solved exactly.

The pioneering work on ISAM uses an approximation that is only valid for low NA systems with narrowband illumination [99]. The approximated inverse problem can be solved in exact form using efficient numerical methods.

The primary goal of ISAM is to correct depth-dependent defocusing effects. This correction is especially important as the NA increases. In this sense, the requirement of low NA in Section 3.4.1 is unsatisfying—the approximation fails in the large NA regime, where we expect to gain the most. Later work used more sophisticated approximations that are valid for high NA and broadband systems [100, 102]. The resulting inverse problem can be solved approximately using the same efficient numerical methods.

While the latter approach is more general, our work in Chapter 4 is closely related to the low NA, narrowband formulation. Next, we describe these two varieties of approximate ISAM formulations.

3.4.1 ISAM Kernel: Low NA and Narrowband Illumination

First, it is assumed that the system has a low numerical aperture. In this case, the Gaussian functions $\hat{g}(\mathbf{k}_\parallel, k_0)$ in (3.4) decay quickly in $\mathbf{k}_\parallel$, and the integrand is effectively zero unless $\mathbf{k}_\parallel$ and $\mathbf{k}'_\parallel$ are much less than k_a. In this regime, we can invoke the *paraxial approximation*, and replace the complex phase terms in (3.4) by the quadratic approximation

$$k_z\left(\mathbf{k}_\parallel, k_0\right) \approx k_0 - \frac{\left|\mathbf{k}_\parallel\right|^2}{2k_0}. \tag{3.5}$$

It is further assumed that the bandwidth of the illuminating source is much lower than its central wavenumber, *i.e.*

$$|k_b - k_a| \ll \frac{k_b + k_a}{2} \triangleq \mu.$$

This is the so-called *narrowband approximation.*

After invoking the paraxial and narrowband approximations, the ISAM kernel has a simple form:

$$\hat{A}_p(\mathbf{k}_\parallel, z - z_F, k_0) \triangleq H_p(\mathbf{k}_\parallel, k_0)\zeta_p(z - z_F)e^{\mathrm{i}\phi_p(\mathbf{k}_\parallel, 2k_0)(z - z_F)},$$

where

$$\begin{aligned}
\chi\left(\mathbf{k}_\parallel, k_0\right) &\triangleq \begin{cases} 1, & \left|\mathbf{k}_\parallel\right| \le k_0 \ \& \ k_a \le k_0 \le k_b \\ 0, & \text{otherwise,} \end{cases} \\
H_p(\mathbf{k}_\parallel, k_0) &\triangleq k_0 \hat{g}\left(\frac{\mathbf{k}_\parallel}{2}, k_0\right)^2 \chi\left(\mathbf{k}_\parallel, k_0\right) \\
\zeta_p(z) &\triangleq \left(1 + \mathrm{i}\frac{\mathrm{NA}^2}{2}\mu z\right)^{-1} \\
\phi_p\left(\mathbf{k}_\parallel, k_0\right) &\triangleq \left(k_0 - \frac{\left|\mathbf{k}_\parallel\right|^2}{2k_0}\right),
\end{aligned}$$

and the subscript p denotes use of the paraxial approximation.

3.4.2 ISAM Kernel: Asymptotic Approximations

The low-NA requirement can be replaced by a pair of asymptotic approximations. We present only the relevant results; for details, see [100].

The first approximation is valid when $|z - z_F| \ll 1/(k_0\mathrm{NA}^2)$. Here, the complex phase term of the integrand in (3.4) is slowly varying, and the integrand is dominated by the product $\hat{g}\left(\mathbf{k}'_\parallel, k_0\right)\hat{g}\left(\mathbf{k}_\parallel - \mathbf{k}'_\parallel, k_0\right)$. As $\hat{g}$ is isotropic in its first argument, this product is concentrated around $\mathbf{k}'_\parallel \approx \mathbf{k}_\parallel/2$. The complex phase can be linearized about this point to obtain a tractable integral. Ultimately, the approximate kernel in the near-focus region is

$$\hat{A}_n(\mathbf{k}_\parallel, z, k_0) = \frac{\pi}{k_z\left(\mathbf{k}_\parallel, 2k_0\right)}\hat{g}\left(\frac{\mathbf{k}_\parallel}{2}, k_0\right)^2 e^{\mathrm{i}k_z(\mathbf{k}_\parallel, 2k_0)z}.$$

The second approximation is valid when $|z - z_F| \gg 1/(k_0\mathrm{NA}^2)$. In this region, the integrand is highly oscillatory and can be estimated using the method of stationary phase. To first order, the asymptotic behavior of $\hat{A}(\mathbf{k}_\parallel, z - z_F, k_0)$ is determined by the value of the integrand at the critical points of the phase; *i.e.*, locations where

$$\nabla_{\mathbf{k}_\parallel}\left(k_z\left(\mathbf{k}'_\parallel, k_0\right) + k_z\left(\mathbf{k}_\parallel - \mathbf{k}'_\parallel, k_0\right)\right) = 0.$$

In a fortuitous turn of events, the critical points are $\mathbf{k}'_\parallel = \mathbf{k}_\parallel/2$; the same as the linearization point in the near-focus case. The approximate kernel in the far-from-focus region is

$$\hat{A}_f(\mathbf{k}_\parallel, z, k_0) = \frac{\pi}{4} e^{\mathrm{i}\frac{\pi}{4}} \frac{k_z\left(\mathbf{k}_\parallel, 2k_0\right)}{k_0} \frac{1}{|z|} \hat{g}\left(\frac{\mathbf{k}_\parallel}{2}, k_0\right)^2 e^{\mathrm{i}k_z(\mathbf{k}_\parallel, 2k_0)z}.$$

Note the similarity between the near-focus and far-from-focus approximations; in particular, the exponential term is the same in both regions. To make the connection explicit, define the phase function

$$\phi_a\left(\mathbf{k}_\parallel, k_0\right) \triangleq k_z\left(\mathbf{k}_\parallel, 2k_0\right),$$

where we use the subscript a to associate ϕ_a with the pair of asymptotic approximations. Define

$$\begin{aligned}
H_n(\mathbf{k}_\parallel, k_0) &\triangleq \frac{\pi}{k_z\left(\mathbf{k}_\parallel, 2k_0\right)} \hat{g}\left(\frac{\mathbf{k}_\parallel}{2}, k_0\right)^2 \chi\left(\mathbf{k}_\parallel, k_0\right) \\
H_f(\mathbf{k}_\parallel, k_0) &\triangleq \frac{\pi}{4} e^{\mathrm{i}\frac{\pi}{4}} \frac{k_z\left(\mathbf{k}_\parallel, 2k_0\right)}{k_0} \hat{g}\left(\frac{\mathbf{k}_\parallel}{2}, k_0\right)^2 \chi\left(\mathbf{k}_\parallel, k_0\right) \\
\zeta_n(z) &\triangleq 1 \\
\zeta_f(z) &\triangleq |z|^{-1}.
\end{aligned}$$

The subscripts n and f indicate “near-focus” and “far-from-focus”, respectively. Now, we can write the approximate ISAM kernels as

$$\begin{aligned}
\hat{A}_n(\mathbf{k}_\parallel, z, k_0) &= H_n(\mathbf{k}_\parallel, k_0)\zeta_n(z) e^{\mathrm{i}\phi_a(\mathbf{k}_\parallel, k_0)z}, \\
\hat{A}_f(\mathbf{k}_\parallel, z, k_0) &= H_f(\mathbf{k}_\parallel, k_0)\zeta_f(z) e^{\mathrm{i}\phi_a(\mathbf{k}_\parallel, k_0)z}.
\end{aligned}$$

Following the approach described in [100], we can form a unified approximation by defining the piecewise functions

$$\zeta_a(z) \triangleq \begin{cases} \zeta_n(z - z_F) & |z - z_F| \ll \frac{1}{k_0 \mathrm{NA}^2} \\ \zeta_f(z - z_F) & |z - z_F| \gg \frac{1}{k_0 \mathrm{NA}^2} \end{cases}$$

$$H_a(\mathbf{k}_\parallel, k_0) \triangleq \begin{cases} H_n(\mathbf{k}_\parallel, k_0) & |z - z_F| \ll \frac{1}{k_0 \mathrm{NA}^2} \\ H_f(\mathbf{k}_\parallel, k_0) & |z - z_F| \gg \frac{1}{k_0 \mathrm{NA}^2}, \end{cases}$$

and writing the approximate kernel as

$$\hat{A}_a(\mathbf{k}_\parallel, z - z_F, k_0) \triangleq H_a(\mathbf{k}_\parallel, k_0)\zeta_a(z - z_F) e^{\mathrm{i}(\phi_a(\mathbf{k}_\parallel, k_0)(z - z_F)}. \qquad (3.6)$$

Strictly speaking, H_a and depends on $z - z_F$ and ζ_a depends on k_0, as these quantities determine the transition between the near-focus and far-from-focus regimes. We will return to this point in

Section 3.6.1.

Note that (3.6) and the narrowband/low-NA approximation (3.5) share the same functional form. While H_a and ζ_a are differ significantly from the paraxial versions H_p and ζ_p, the phase functions are closely related—indeed, ϕ_p can be obtained from ϕ_a by way of the quadratic approximation (3.5).

3.5 Projection-Slice Interpretation

The paraxial and asymptotic approximations result in approximate ISAM kernels with similar forms. For now, we ignore the distinction between the two types of approximate models, and instead assume the ISAM kernel can be approximated in the form

$$\hat{A}(\mathbf{k}_\parallel, z - z_F, k_0) \approx H(\mathbf{k}_\parallel, k_0)\zeta(z - z_F)e^{\mathrm{i}(\phi(\mathbf{k}_\parallel, k_0)(z - z_F)}. \tag{3.7}$$

This form provides tremendous insight into the nature of the ISAM imaging.

Consider a single, fixed focal plane; this is the usual setting for ISAM imaging. For the remainder of this section we move the dependence on the focal plane to a subscript; that is, we have $\hat{S}_{z_F}(\mathbf{k}_\parallel, k_0) = \hat{S}(\mathbf{k}_\parallel, z_F, k_0)$.

Define the *weighted susceptibility*

$$\hat{\xi}_{z_F}(\mathbf{k}_\parallel, z, k_0) \triangleq \zeta(z - z_F)\hat{\eta}(\mathbf{k}_\parallel, z, k_0).$$

Inserting the generic approximate ISAM kernel (3.7) into the measurement equation (3.4) yields

$$\begin{aligned}\hat{S}_{z_F}(\mathbf{k}_\parallel, k_0) &\approx H(\mathbf{k}_\parallel, k_0)\int \zeta(z - z_F)\hat{\eta}(\mathbf{k}_\parallel, z, k_0)e^{\mathrm{i}\phi(\mathbf{k}_\parallel, k_0)(z - z_F)}\mathrm{d}z \\ &= H(\mathbf{k}_\parallel, k_0)\int \hat{\xi}_{z_F}(\mathbf{k}_\parallel, z, k_0)e^{\mathrm{i}\phi(\mathbf{k}_\parallel, k_0)(z - z_F)}\mathrm{d}z \\ &= H(\mathbf{k}_\parallel, k_0)e^{\mathrm{i}\phi(\mathbf{k}_\parallel, k_0)z_F}\hat{\hat{\xi}}_{z_F}\left(\mathbf{k}_\parallel, -\phi\left(\mathbf{k}_\parallel, k_0\right), k_0\right), \end{aligned} \tag{3.8}$$

where the double hat indicates a 3D Fourier transform with respect to $\mathbf{r} = (x, y, z)$.

Equation (3.8) is a generalized projection-slice theorem: the ISAM data is approximately the bandlimited Fourier transform (with respect to $\mathbf{r}$) of the weighted susceptibility evaluated on a three-dimensional surface parameterized by $\mathbf{k}_\parallel$ and k_0. By varying $\mathbf{k}_\parallel$ and k_0, we are able to observe a 3D slice of the four-dimensional function $\hat{\hat{\xi}}_{z_F}(\mathbf{k}_\parallel, k_z, k_0)$ constrained to the surface

$$\mathsf{V} \triangleq \left\{(k_x, k_y, k_z, k_0) : \sqrt{k_x^2 + k_y^2 + k_z^2} = 2k_0,\ k_z < 0,\ k_x^2 + k_y^2 \leq 4(k_0\mathrm{NA})^2,\ k_a \leq k_0 \leq k_b\right\}.$$

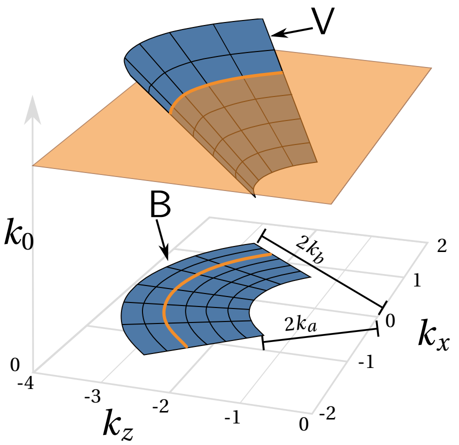


Figure 3.3: Observable Fourier components for a target with two spatial and one spectral dimensions. The intersection of V with a plane of constant k_0 becomes an arc of constant radius when projected onto the (k_x, k_z) plane.

The sampling surface for a target with two spatial dimensions, *i.e.* $\mathbf{r} = (x, z)$, is illustrated in Fig. 3.3; that we can only observe $k_z < 0$ is due to the backscattering geometry. As defined, V contains only the Fourier components above the e^{-2} cutoff frequency of $\hat{g}(\mathbf{k}_\parallel/2, k_0)^2$; this is an arbitrary choice as $\hat{g}$ decays smoothly in $|\mathbf{k}_\parallel|$.

We cannot recover an arbitrary object given ISAM data at a single focal plane. Even analytic continuation is not possible in this setting, as such methods require data over a four-dimensional volume element and we are restricted to a three-dimensional surface [103]. If we were to scan along z_F in addition to $\mathbf{r}_\parallel^{(o)}$, we could further simplify by taking a Fourier transform along z_F. The measurements would be of the form $\hat{\hat{S}}(\mathbf{k}_\parallel, k_z, k_0) = \hat{\hat{A}}(\mathbf{k}_\parallel, k_z, k_0)\hat{\hat{\eta}}(\mathbf{k}_\parallel, k_z, k_0)$, where the double hat indicates the 3D Fourier transform with respect to $\mathbf{r}$. Now, η can be recovered using a standard deconvolution procedure. Unfortunately, this approach is not desirable due to the amount of data required. This problem is the focus of Chapter 5 of this thesis.

The situation is simplified if η is not a function of k_0; such an object is said to be *non-dispersive*. This is one of the key assumptions on which ISAM, optical coherence tomography, diffraction tomography, and reflection tomography are built [104–106]. In this case, the measurements are related to a 3D slice of the 3D target $\eta(x, y, z)$. The observable Fourier components are

$$\mathsf{B} \triangleq \left\{(k_x, k_y, k_z) : \sqrt{k_x^2 + k_y^2 + k_z^2} = 2k_0,\ k_z < 0,\ k_x^2 + k_y^2 \leq 4(k_0 \mathrm{NA})^2,\ k_a \leq k_0 \leq k_b\right\}.$$

The region B is called the *optical passband* of the ISAM imaging system. Strictly speaking, we observe the Fourier components of the *weighted* susceptibility on B, but this distinction is usually ignored; we return to this point in Section 3.6.1. Only a non-dispersive (weighted) object whose spatial Fourier transform is supported on B can be perfectly imaged by the ISAM system with a single focal plane. Otherwise, ISAM is able to recover, at best, a spatial bandpass version of the original target. In the visualization of Fig. 3.3, B is the "shadow" of V on the plane $k_0 = 0$.

3.6 Filtered Back Projection for ISAM

We restrict our attention to the case of a non-dispersive object and data acquired from a single focal plane; this is the usual setting for ISAM imaging. In this case, the exact ISAM measurement equation (3.3) reduces to

$$\hat{S}_{z_F}(\mathbf{k}_\parallel, k_0) = \int \hat{A}(\mathbf{k}_\parallel, z - z_F, k_0)\hat{\eta}(\mathbf{k}_\parallel, z)\mathrm{d}z. \tag{3.9}$$

Observe that (3.9) is a "diagonal" relationship in $\mathbf{k}_\parallel$, as can be expected due to the transverse shift invariance property of the imaging system. For each $\mathbf{k}_\parallel$, (3.9) is a Fredholm integral equation of the first kind and can be solved with standard numerical methods; *e.g.* the method of moments, the Galerkin method, and the projection method, among others [107, 108].

Solving the perturbed ISAM problem using the approximations described in Sections 3.4.1 and 3.4.2 leads to fast reconstruction algorithms. In the non-dispersive case, the projection-slice relationship (3.8) reduces to

$$\hat{S}_{z_F}(\mathbf{k}_\parallel, k_0) = H(\mathbf{k}_\parallel, k_0)e^{\mathrm{i}\phi(\mathbf{k}_\parallel, k_0)z_F}\hat{\hat{\xi}}_{z_F}\left(\mathbf{k}_\parallel, -\phi\left(\mathbf{k}_\parallel, k_0\right)\right). \tag{3.10}$$

Stated again, the measurements are the filtered and bandlimited 3D Fourier transform of the weighted susceptibility evaluated at coordinates $(\mathbf{k}_\parallel, \phi(\mathbf{k}_\parallel, k_0))$. This suggests a straightforward reconstruction technique: simply undo the filtering, take an appropriate inverse Fourier transform, and then unweight the reconstructed weighted susceptibility. This reconstruction algorithm is similar to the generalized filtered backprojection algorithm developed for diffraction tomography [106, 109].

As the ISAM imaging operator has a large nullspace, we must apply some form of regularization. In our case, we must apply a regularized inverse to the linear operator $\mathcal{H} : (\mathcal{H}f)(\mathbf{k}_\parallel, k_0) = H(\mathbf{k}_\parallel, k_0)f(\mathbf{k}_\parallel, k_0)$. One option is to use a Tikhonov regularized inverse. We introduce the scalar regularization parameter $\lambda > 0$ and set

$$(\mathcal{H}^\dagger f)(\mathbf{k}_\parallel, k_0) = \frac{f(\mathbf{k}_\parallel, k_0)}{H(\mathbf{k}_\parallel, k_0) + \lambda}.$$

Applying the Tikhonov regularized filter, we have

$$\begin{aligned}\hat{S}'_{z_F}(\mathbf{k}_\parallel, k_0) &= \frac{\hat{S}_{z_F}(\mathbf{k}_\parallel, k_0)}{\lambda + H(\mathbf{k}_\parallel, k_0)} e^{-\mathrm{i}\phi(\mathbf{k}_\parallel, k_0) z_F} \\ &= \frac{H(\mathbf{k}_\parallel, k_0)}{\lambda + H(\mathbf{k}_\parallel, k_0)} \hat{\hat{\xi}}_{z_F}\left(\mathbf{k}_\parallel, -\phi\left(\mathbf{k}_\parallel, k_0\right)\right).\end{aligned}$$

Next, we must take an inverse Fourier transform. In the usual ISAM imaging scenario, we acquire discrete data that is uniformly sampled in $\mathbf{k}_\parallel$ and k_0. To apply fast numerical methods, we first resample the data to be uniformly spaced in $\mathbf{k}_\parallel$ and $k_z = \phi(\mathbf{k}_\parallel, k_0)$ and then apply a 3D inverse Fast Fourier Transform (FFT) with respect to $\mathbf{k}_\parallel$ and k_z.

For a fixed value of $\mathbf{k}_\parallel$ and viewed as a univariate function of k_0, ϕ is one-to-one for $k_0 \in [k_a, k_b]$, and thus invertible on this interval. We must resample the data at an evenly spaced grid of k_z according to $k_0 = \phi^{-1}(\mathbf{k}_\parallel, k_z)$. If we are using the asymptotic approximations described in Section 3.4.2, then $\phi(\mathbf{k}_\parallel, k_0) = \sqrt{4k_0^2 - \left|\mathbf{k}_\parallel\right|^2}$ and we must apply the coordinate transformation

$$k_0 \mapsto \frac{1}{2}\sqrt{k_z^2 + \left|\mathbf{k}_\parallel\right|^2}.$$

We have

$$\begin{aligned}\hat{S}''_{z_F}(\mathbf{k}_\parallel, k_z) &= \hat{S}'_{z_F}(\mathbf{k}_\parallel, \phi^{-1}(\mathbf{k}_\parallel, k_z)) \\ &= \frac{H(\mathbf{k}_\parallel, \phi^{-1}(\mathbf{k}_\parallel, k_z))}{\lambda + H(\mathbf{k}_\parallel, \phi^{-1}(\mathbf{k}_\parallel, k_z))} \hat{\hat{\xi}}_{z_F}\left(\mathbf{k}_\parallel, k_z\right).\end{aligned}$$

Note that this resampling operation arises in a variety of synthetic aperture tomographic imaging modalities: in seismic imaging it is known the Stolt mapping [110], in Synthetic Aperture Radar (SAR) it is known as the $\omega - k$ algorithm [111, 112], and also arises as a step in the synthetic aperture focusing technique in ultrasound imaging [113]. Alternatives to the use of interpolation and inverse FFT have been developed for these similar imaging problems. Algorithms based on gridding or the non-uniform FFT provide user-controllable accuracy bounds. Interpolation-free approaches have also been proposed. See, *e.g.*, [114–116] and the references within.

After taking the inverse Fourier transform of the resampled data, we have a filtered version of the weighted susceptibility

$$S''(\mathbf{r}) = (h * \xi_{z_F})(\mathbf{r}),$$

where $*$ denotes three-dimensional convolution with respect to $\mathbf{r}$ and the filter h is given by

$$h(\mathbf{r}) = \int \chi_{\mathrm{B}}(\mathbf{k}) \, \frac{H(\mathbf{k}_\parallel, \phi^{-1}(\mathbf{k}_\parallel, k_z))}{\lambda + H(\mathbf{k}_\parallel, \phi^{-1}(\mathbf{k}_\parallel, k_z))} e^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} \mathrm{d}^3 k$$

Algorithm 3 Fourier Inversion for ISAM

```
INPUT: ISAM data $\hat{S}_{z_F}(\mathbf{k}_\parallel, k_0)$; Regularization parameter $\lambda$
OUTPUT: Object estimate $\tilde{\eta}(\mathbf{r})$
$\hat{S}'_{z_F}(\mathbf{k}_\parallel, k_0) = \hat{S}_{z_F}(\mathbf{k}_\parallel, k_0) / \left(H(\mathbf{k}_\parallel, k_0) + \lambda\right)$    ▷ Apply (regularized) inverse filter
$\hat{S}''_{z_F}(\mathbf{k}_\parallel, k_z) = \hat{S}_{z_F}\left(\mathbf{k}_\parallel, k_0 = \phi^{-1}(\mathbf{k}_\parallel, k_z)\right)$    ▷ Resample
$S''_{z_F}(\mathbf{r}) = \tilde{\eta}(\mathbf{r})\zeta(z - z_F)$    ▷ 3D Inverse FFT
$\tilde{\eta}(\mathbf{r}) = \tilde{\eta}(\mathbf{r}) / \zeta(z - z_F)$    ▷ Compensate for signal decay
```

and χ_{B} denotes the indicator function for the optical passband.

The resolution of the ISAM system depends on the size of the optical passband. In realistic imaging scenarios the filter h varies much faster than the slowly varying weighting function, ζ. If the object consists of small, well-separated scatterers, it is reasonable to make the approximation

$$(h * \xi_{z_F})(\mathbf{r}) = \left(h * (\zeta(z)\eta(\mathbf{r}))\right)(\mathbf{r}) \approx (h * \eta)(\mathbf{r})\zeta(z).$$

Finally, we can remove the effect of $\zeta(z)$ by simple division, or by performing a regularized inversion such as $(\zeta(z) + \tau)^{-1}$ for some $\tau > 0$.

This Fourier inversion algorithm is summarized in Algorithm 3.

3.6.1 Remarks

The ISAM Fourier inversion algorithm relies on approximating the ISAM kernel in the form (3.10). The key aspect is that the variables k_0 and z interact only in the complex exponential term.

It is helpful to frame this in terms of an operator factorization. Define the *ISAM operator* $\mathcal{A} : L^2(\mathbb{R}^3) \to L^2(\mathbb{R}^3)$ which uses the exact ISAM kernel;

$$(\mathcal{A}_{z_F}\eta)(\mathbf{k}_\parallel, k_0) \triangleq \int_{-\infty}^{\infty} \hat{A}(\mathbf{k}_\parallel, z - z_F, k_0)\eta(\mathbf{k}_\parallel, z)\mathrm{d}z,$$

as well as the *approximate ISAM operator* $\tilde{\mathcal{A}}_{z_F}$ which uses (3.10) as its integration kernel;

$$(\tilde{\mathcal{A}}_{z_F}\eta)(\mathbf{k}_\parallel, k_0) \triangleq \int H(\mathbf{k}_\parallel, k_0)\zeta(z - z_F)e^{\mathrm{i}(\phi(\mathbf{k}_\parallel, k_0)(z - z_F)}\eta(\mathbf{k}_\parallel, z)\mathrm{d}z.$$

Additionally, define the operators $\mathcal{F}_{z_F}, \mathcal{H}, \mathcal{V}_{z_F} : L^2(\mathbb{R}^3) \to L^2(\mathbb{R}^3)$

$$
\begin{aligned}
(\mathcal{F}_{z_F}\eta)(\mathbf{k}_\parallel, k_0) &\triangleq (2\pi)^{-\frac{3}{2}} \int_{-\infty}^{\infty} e^{-\mathrm{i}\phi(\mathbf{k}_\parallel, k_0)(z - z_F)} \eta(\mathbf{k}_\parallel, z)\,\mathrm{d}z, \\
(\mathcal{H} f)(\mathbf{k}_\parallel, k_0) &\triangleq (2\pi)^{\frac{3}{2}} H(\mathbf{k}_\parallel, k_0) f(\mathbf{k}_\parallel, k_0), \\
(\mathcal{V}_{z_F}\eta)(\mathbf{k}_\parallel, z) &\triangleq \zeta(z - z_F)\eta(\mathbf{k}_\parallel, z).
\end{aligned}
$$

Now, $\tilde{A}_{z_F} = \mathcal{H}\mathcal{F}_{z_F}\mathcal{V}_{z_F}$. Each of these operators is easy to (pseudo)invert—$\mathcal{H}$ and $\mathcal{V}_{z_F}$ require only division, and $\mathcal{F}_{z_F}$ is unitary.

This factorization follows directly the approximations used in the narrowband and paraxial regime. However, there are difficulties when using the asymptotic approximations described in Section 3.4.2. In this case, H_a and ζ_a depend on both k_0 and z, as the transition between the near-focus and far-from-focus approximations depends on $|z - z_F|\, k_0 \mathrm{NA}^2$. We can reconcile this difficulty in a few ways. One option is to redefine H_a as the average of H_n and H_f, and to define

$$
\zeta_a(z) = \begin{cases} \zeta_n(z - z_F) & |z - z_F| \ll \frac{1}{\mu \mathrm{NA}^2} \\ \zeta_f(z - z_F) & |z - z_F| \gg \frac{1}{\mu \mathrm{NA}^2}, \end{cases} \tag{3.11}
$$

where $\mu = (k_b + k_a)/2$ is the central illuminating wavenumber. This approach is suitable for reconstruction using Algorithm 3. However, it results in catastrophic reconstruction errors when used in model-based iterative reconstructions, as we will discuss Section 3.7. This challenge motivates the low-rank approximation approach we present in Chapter 4.

There is a final option to reconcile the effects of H_a and ζ_a: simply ignore them. As pointed out by the authors in several publications, the resampling from k_0 to k_z is the most important step in the Fourier inversion algorithm. It is this step that compensates for defocusing away from the focal plane, and is thus responsible for the property of depth-invariant resolution.

Note that traditional OCT reconstruction is based on the inverse Fourier transform of (3.10); that is, the effect of $\phi(\mathbf{k}_\parallel, k_0)$ is ignored. This is the root cause of defocusing effects, and is why OCT is traditionally limited to low-NA systems.

The effect of each step in the ISAM Fourier inversion algorithm is illustrated in Fig. 3.4; horizontal and vertical profiles through the reconstructions are shown in Fig. 3.5. The OCT reconstruction shows a clear loss of transverse resolution away from the focal plane; this is expected, as the mean confocal parameter for this system is ≈ 2.7 µm. Figure 3.4(b) is the reconstruction using only the resampling step. The reconstruction exhibits the desired depth-invariant transverse spatial resolution, however the loss of signal intensity away from the focal plane is evident. Figure 3.4(c) is the reconstruction with $\mathcal{H}^{-1}$. Increasing the regularization

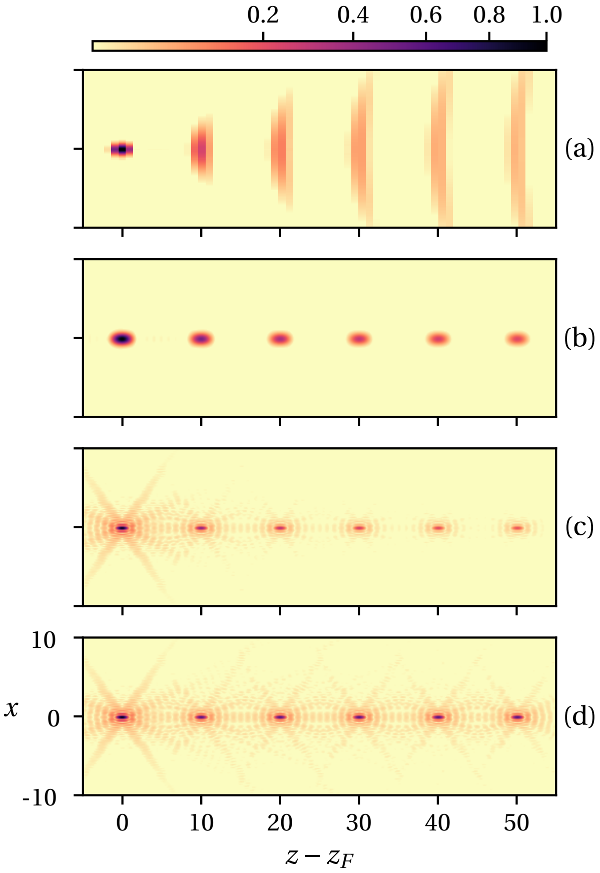


Figure 3.4: All units µm. Comparing reconstruction methods when applied to point targets. Point scatterers are located at $z - z_F = 0, 10, \ldots, 50$ µm. System NA = 0.3; $k_a = 6.6, k_b = 10$ rad^{-1}. (a) OCT reconstruction exhibits strong defocusing artifacts. Transverse resolution is lost away from the focal region. (b) Reconstruction using resampling only. Transverse resolution is uniform for all depths. (c) Inverting the effect of $H(\mathbf{k}_\parallel, k_0)$ further improves transverse resolution. (d) Compensating for beam decay away from focus.

parameter results in a loss of spatial resolution, but also reduces ringing. Finally, Fig. 3.4(d) shows the effect of applying $\mathcal{V}_{n_F}^{-1}$; here, we used the hybrid weighting function (3.11). The 1D profile taken at $x = 0$ is shown in Fig. 3.5. Note the difference in recovered scatterer amplitude. Only the scatterer at $z - z_F = 0$ is in the near-focus regime.

3.7 Sampling, Discretization, and Reconstruction

Thus far, we have discussed the ISAM inverse problem in the continuous setting. To adopt the language of Myers and Barrett, this is the *continuous to continuous* (CC) setting: the object and

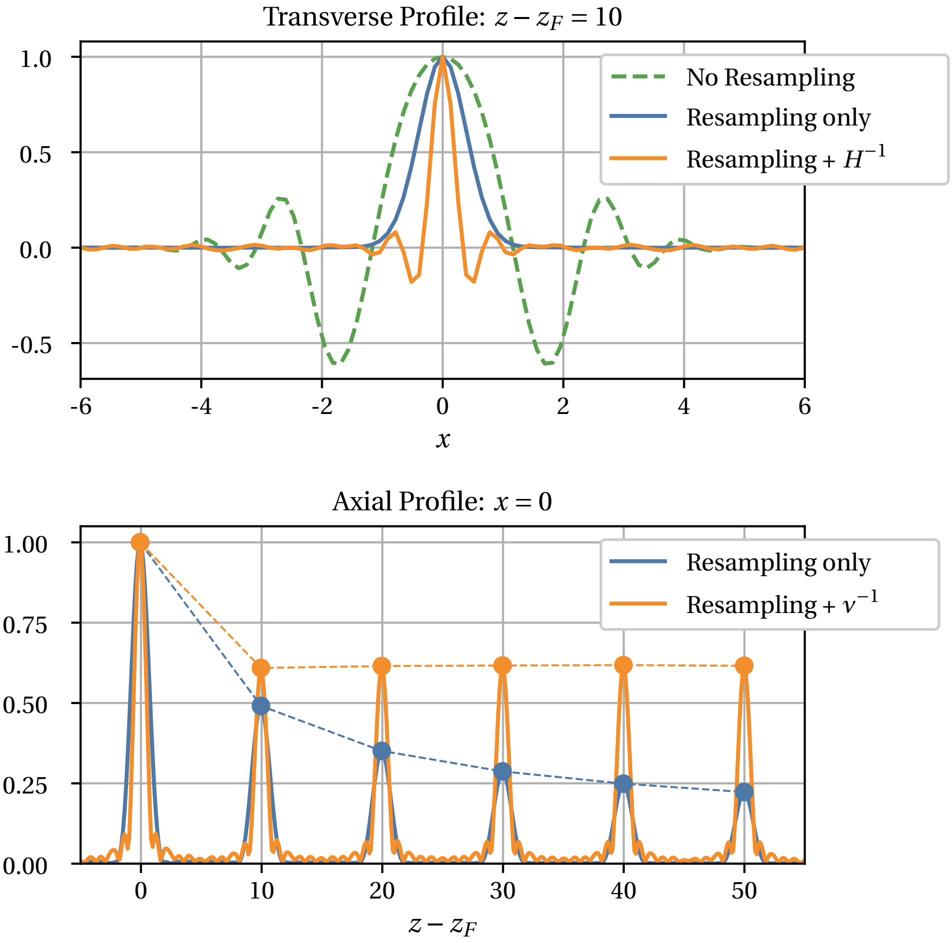


Figure 3.5: One-dimensional profiles of the magnitude of reconstructed objects shown in Fig. 3.4. Each profile was normalized to have unit magnitude at $(x, z) = (0, z_F)$.

data are viewed as functions, and the data is related to the object via an integral operator [117].

Of course, in practical imaging problems we do not acquire a continuum of data. Data is only measured at a finite collection of points. There are a variety of ways to attack such *continuous to discrete* (CD) inverse problems. See, *e.g.*, [117–120] for an overview of these methods. A common approach is to solve the (possibly regularized) CC problem and to use numerical methods to discretize the solution. An immediate example is the application of the Fourier inversion algorithm to sampled data.

In portions of Chapter 4, and all of Chapter 5, we use a different approach to solve the ISAM inverse problem (3.9). We project the inverse problem into a finite dimensional space and use numerical methods to solve the finite dimensional problem; this is sometimes called *the projection method* [107, 117]. This approach does not use the approximations discussed in Section 3.4; instead, the unapproximated ISAM kernel (3.4) is used in the forward model (3.9).

Next, we discuss the problems of sampling, discretization, and variational image reconstruction.

3.7.1 Sampling

The instrument acquires samples of the spatial-domain measurement equation (3.2). We assume the object is (spatially) supported in a region $\Gamma \subset \mathbb{R}^3$; here, we take $\Gamma = [0, L_x] \times [0, L_y] \times [0, L_z]$. We write the number of samples as N_i and the discretization or sampling interval as Δ_i for $i = x, y, z, k$. We obtain measurements at the transverse aperture locations $\mathbf{r}_\parallel^{(o)} = (n_x \Delta_x, n_y \Delta_y)$ for integers n_x, n_y. The parameters are chosen to cover Γ, *i.e.* $N_i \Delta_i = L_i$ holds for $i = x, y, z$. For simplicity, we assume the sampling parameters are the same along the x and y directions: $N_x = N_y$, $\Delta_x = \Delta_y$, and $L_x = L_y = N_x \Delta_x$. The wavenumber is sampled uniformly over the interval $[k_a, k_b]$ with sampling interval Δ_k; the n_k-th measurement wavenumber is $k_{0,i} \triangleq k_a + n_k \Delta_k$. We acquire data at N_F focal planes, written $\{z_{F,i} : i = 1, 2, \dots N_F\}$, where $N_F = 1$ for the standard ISAM problem. The same sampling parameters are used at each focal plane; in particular, the set of sampled wavenumbers does not change.

We choose the sampling parameters as we would for a standard, single-species ISAM problem. The necessary sampling intervals can be motivated using the approximate forward models. Using these models model, it can be shown that "point spread function" $\left|A(\mathbf{r}_\parallel, k_0, z)\right|$ (approximately) decays like a Gaussian in $\left|\mathbf{r}_\parallel\right|$. We take L_x and L_y large enough to safely neglect the unmeasured data. Moreover, for fixed z_F the measurements $\hat{S}(\mathbf{k}_\parallel, k_0, z_F)$ are bandlimited to $[-k_b \sin \mathrm{NA}, k_b \sin \mathrm{NA}]$; we sample along the transverse dimension at intervals $\Delta_x, \Delta_y < \pi / (k_b \sin \mathrm{NA})$. Finally, the combination of uniform sampling in $\mathbf{r}_\parallel^{(o)}$ and k_0 leads to a non-uniform sampling of the Fourier transform of the object: samples are obtained at uniform locations along the $\mathbf{k}_\parallel$ axis but at nonuniform locations along the k_z axis. To avoid aliasing, we require that the maximum distance between samples on the k_z axis is less than π / L_z [121, 122].

3.7.2 Discretization and Block-Diagonal Matrix Structure

Given $N_x \times N_y$ spatial samples and N_k wavenumber samples of (3.2), we take the 2D Discrete Fourier Transform (DFT) with respect to the transverse coordinates and write the result as the tensor $\hat{\mathbf{S}} \in \mathbb{C}^{N_x \times N_y \times N_k \times N_F}$. We continue to assume $N_x = N_y$ with N_x an even integer. The 2D-DFT coordinate $\mathbf{q}_\parallel = (q_x, q_y)$ is an integer vector with $0 \le q_x, q_y \le N_x - 1$. We obtain the continuous

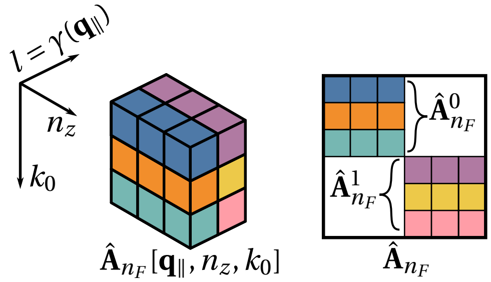

$l=\gamma(\mathbf{q}_\parallel)$
n_z
k_0
$\hat{\mathbf{A}}_{n_F}[\mathbf{q}_\parallel, n_z, k_0]$
$\hat{\mathbf{A}}^0_{n_F}$
$\hat{\mathbf{A}}^1_{n_F}$
$\hat{\mathbf{A}}_{n_F}$

Figure 3.6: Unfolding the tensor $\hat{\mathbf{A}}_{n_F}[\mathbf{q}_\parallel, n_z, n_k]$ into a block-diagonal matrix.

Fourier coordinate k_x from the DFT coordinate k_x as

$$k_x(q_x) = \begin{cases} 2\pi q_x/L_x & q_x < N_x/2 \\ 2\pi(q_x - N_x)/L_x & \text{otherwise,} \end{cases}$$

and the same holds for q_y and k_y. We define $\mathbf{k}_\parallel(\mathbf{q}_\parallel) = \big(k_x(q_x), k_y(q_y)\big)$. We use the reindexing function

$$\gamma : \mathbb{Z}^2 \to \mathbb{Z} \quad \gamma(\mathbf{q}_\parallel) = q_x + N_x q_y,$$

to identify the 2D-DFT index $\mathbf{q}_\parallel$ with the integer $\gamma(\mathbf{q}_\parallel)$.

The discretized ISAM measurement model is

$$\hat{\mathbf{S}}[\mathbf{q}_\parallel, n_k, n_F] = \sum_{n_z=0}^{N_z-1} \hat{\mathbf{A}}_{n_F}[\mathbf{q}_\parallel, n_k, n_z]\hat{\boldsymbol{\eta}}[\mathbf{q}_\parallel, n_z] \tag{3.12}$$

$\hat{\boldsymbol{\eta}} \in \mathbb{C}^{N_x \times N_y \times N_z}$ is the discretized (non-dispersive) susceptibility to be imaged, and

$$\hat{\mathbf{A}}_{n_F}[\mathbf{q}_\parallel, n_k, n_z] \triangleq \hat{A}(\mathbf{k}_\parallel(\mathbf{q}_\parallel), k_a + n_k\Delta_k, N_z\Delta_z - z_{F,n_F}).$$

We can write (3.12) as a matrix-vector product with a certain block structure. We use a superscript to denote a submatrix or vector formed for particular value of $\mathbf{q}_\parallel$. For each $n_F \in [N_F]$ and $l = \gamma(\mathbf{q}_\parallel)$, define $\hat{\mathbf{A}}^l_{n_F} \in \mathbb{C}^{N_k \times N_z}$ and $\hat{\boldsymbol{\eta}}^l \in \mathbb{C}^{N_z}$ by

$$\begin{aligned} \hat{\mathbf{A}}^l_{n_F} &= \hat{\mathbf{A}}_{n_F}[\gamma^{-1}(l) :, :] \\ \hat{\boldsymbol{\eta}}^l &= \hat{\boldsymbol{\eta}}[\gamma^{-1}(l), :]. \end{aligned}$$

Now, we have

$$\hat{\mathbf{S}}[\gamma^{-1}(l), : , n_F] = \hat{\mathbf{A}}^l\hat{\boldsymbol{\eta}}^l.$$

Let $\hat{\mathbf{s}}_{n_F} \in \mathbb{C}^{N_x N_y N_k}$ be the vector of measurements acquired at the n_F-th focal plane,

$$\hat{\mathbf{s}}_{n_F} = \mathrm{vec}\left(\hat{\mathbf{S}}[:\,,:\,,n_F]\right),$$

and define

$$\hat{\mathbf{A}}_{n_F} \triangleq \mathrm{blkdiag}\left(\left\{\hat{\mathbf{A}}_{n_F}^{l}\right\}_{l=0}^{N_x N_y - 1}\right) \in \mathbb{C}^{N_x N_y N_k \times N_x N_y N_z}.$$

This construction is illustrated in Fig. 3.6. In an abuse of notation, we use $\hat{\boldsymbol{\eta}}$ to denote the vectorized version of the tensor $\hat{\boldsymbol{\eta}}[\mathbf{q}_{\|}, n_z]$; equivalently, it is the stacking of the $\{\hat{\boldsymbol{\eta}}^l\}$ into a single vector

$$\hat{\boldsymbol{\eta}} = \mathrm{vec}\left(\hat{\boldsymbol{\eta}}[:,:]\right) = \begin{bmatrix} \hat{\boldsymbol{\eta}}^1 \\ \vdots \\ \hat{\boldsymbol{\eta}}^{N_x N_y - 1} \end{bmatrix}.$$

With this definition in place, the (vectorized) measurements from the n_F-th focal plane is

$$\hat{\mathbf{s}}_{n_F} = \hat{\mathbf{A}}_{n_F} \hat{\boldsymbol{\eta}}. \tag{3.13}$$

The matrix-vector product (3.13) is the discretized form of (3.3), assuming η is non-dispersive. The diagonal nature of $\mathcal{A}$ when expressed in the transverse Fourier domain is manifest in the block-diagonal structure of $\hat{\mathbf{A}}_{n_F}$. Finally, we can incorporate data from multiple focal planes by stacking the appropriate data vectors and matrices. In particular, we define the stacked measurement vectors

$$\bar{\mathbf{s}} = \begin{bmatrix} \hat{\mathbf{s}}_1 \\ \vdots \\ \hat{\mathbf{s}}_{N_F} \end{bmatrix} \in \mathbb{C}^{N_x N_y N_k N_F}$$

and the stack of ISAM matrices as

$$\bar{\mathbf{A}} = \begin{bmatrix} \hat{\mathbf{A}}_1 \\ \vdots \\ \hat{\mathbf{A}}_{N_F} \end{bmatrix} \in \mathbb{C}^{N_x N_y N_k N_F \times N_x N_y N_z}.$$

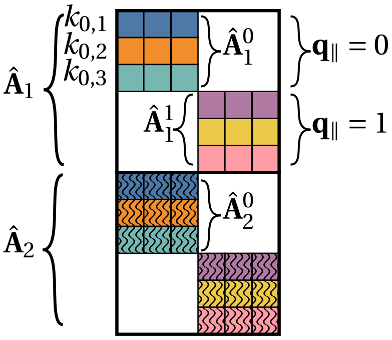


Figure 3.7: Forming $\bar{\mathbf{A}}$ by stacking the $\hat{\mathbf{A}}_{n_F}$. Here, shaded rows correspond to the second focal plane.

Now can write the multifocal ISAM measurement equation

$$\bar{\mathbf{s}} = \bar{\mathbf{A}}\hat{\boldsymbol{\eta}}.$$

The stacking procedure to generate $\bar{\mathbf{A}}$ is illustrated in Fig. 3.7.

3.7.3 Model-Based Reconstruction

We conclude this chapter by discussing the use of optimization-based approaches to solve the discretized ISAM inverse problem. Our task is to recover the (discretized) susceptibility $\hat{\boldsymbol{\eta}}$ from measurements $\bar{\mathbf{s}}$ acquired at one or more focal planes. We begin by assuming a single focal plane, $N_F = 1$; we drop the n_F subscript of 1 and write our forward model as $\hat{\mathbf{s}} = \hat{\mathbf{A}}\hat{\boldsymbol{\eta}}$.

We pose the ISAM reconstruction problem as the variational minimization problem

$$\min_{\hat{\boldsymbol{\eta}}} L(\hat{\mathbf{A}}\hat{\boldsymbol{\eta}}, \hat{\mathbf{s}}) + R(\hat{\boldsymbol{\eta}}),$$

where $L : \mathbb{C}^{N_x N_y N_k} \times \mathbb{C}^{N_x N_y N_k} \to \mathbb{R}$ is a measures the discrepancy between the data, $\hat{\mathbf{s}}$, and the image of a candidate object $\hat{\boldsymbol{\eta}}$ under the forward model $\hat{\mathbf{A}}$. The functional L is often called a *data fidelity* penalty. If we adopt a statistical viewpoint, L is the negative log likelihood of the data, $\hat{\mathbf{s}}$. Thus we can incorporate knowledge of the data statistics into the reconstruction process, and account for effects such as additive noise, shot noise, background signal, and more [123]. This often comes at the cost of a difficult, nonlinear optimization problem [123–125]. A simpler choice, and one that may lead to tractable optimization problems, is to choose

$$L(\hat{\mathbf{A}}\hat{\boldsymbol{\eta}}, \hat{\mathbf{s}}) \triangleq \frac{1}{2}\|\hat{\mathbf{s}} - \hat{\mathbf{A}}\hat{\boldsymbol{\eta}}\|_2^2,$$

leading to the penalized least squares problem

$$\min_{\hat{\boldsymbol{\eta}}} \frac{1}{2} \|\hat{\mathbf{s}} - \hat{\mathbf{A}}\hat{\boldsymbol{\eta}}\|_2^2 + \lambda R(\hat{\boldsymbol{\eta}}). \tag{3.14}$$

The functional $R : \mathbb{C}^{N_x N_y N_z} \to \mathbb{R}$ regularizes the inverse problem and encodes any constraints or *a priori* assumptions regarding $\hat{\boldsymbol{\eta}}$. Tikhonov regularization corresponds to $R(\hat{\boldsymbol{\eta}}) = \|\hat{\boldsymbol{\eta}}\|_2^2$. Alternatively, solutions that are sparse in a transform domain are obtained by setting $R(\boldsymbol{\eta}) = \|\mathbf{C}\boldsymbol{\eta}\|_1$, where $\mathbf{C}$ is a sparsifying transform, *e.g.* finite differences or the wavelet transform. The non-negative scalar λ balances the influence of the data fidelity and regularization terms.

In most cases, the optimization problem (3.14) cannot be solved in closed form and an iterative solution is required. Tikhonov regularization is a notable exception. In this case, (3.14) reduces to the solution of the linear system

$$(\hat{\mathbf{A}}^{\mathsf{H}}\hat{\mathbf{A}} + \lambda\mathbf{I})\hat{\boldsymbol{\eta}} = \hat{\mathbf{A}}^{\mathsf{H}}\hat{\mathbf{s}}, \tag{3.15}$$

where $\mathbf{I}$ is the $N_k N_x N_y \times N_k N_x N_y$ identity matrix. Due to the special structure of $\hat{\mathbf{A}}$, this can be (approximately) solved in closed form using a discretized version of the Fourier inversion algorithm.

Later, we arrive at variations of the ISAM problem for which the equivalent of (3.15) cannot be solved in closed form. In these cases, an iterative solution is required as the Gram matrix $\hat{\mathbf{A}}^{\mathsf{H}}\hat{\mathbf{A}} \in \mathbb{C}^{N_k N_x N_y \times N_k N_x N_y}$ is too large to store, much less invert. The conjugate gradient (CG) algorithm works well in practice. CG requires only matrix-vector products with $\hat{\mathbf{A}}^{\mathsf{H}}$ and $\hat{\mathbf{A}}$. For these methods to be successful, the implementation of $\hat{\mathbf{A}}^{\mathsf{H}}$ and $\hat{\mathbf{A}}$ must be accurate.

Many sparsity-promoting regularizers are non-differentiable. In this case, proximal methods such as FISTA [89] or the Alternating Direction Method of Multipliers (ADMM) [126–128] are attractive. This class of algorithms decomposes the problem (3.14) into a sequence of simpler subproblems. The solution of a linear system similar to (3.15) is often a key ingredient of such algorithms.

Chapter 4

Non-Asymptotic ISAM

4.1 Introduction

We begin by showing that, under the paraxial (low NA) approximation, the ISAM kernel can be written as the product of two terms: an oscillatory term with phase that is linear in z, and a second term that is smooth and well-approximated as the sum of a small number of separable functions. If only a single term is retained, our approach coincides with earlier work for ISAM in the narrowband and paraxial regime. The higher-order terms become appreciable as the bandwidth increases and the narrowband approximation is violated. The separated approximation is the starting point for our two main contributions.

First, we study the spectrum of a perturbed ISAM operator using our factored kernel. We show that the left singular functions of this operator satisfy a certain Sturm-Liouville differential equation.

Second, we use our approximate kernel to develop fast and accurate numerical methods to compute the action of the ISAM operator and its adjoint. Our proposed algorithm requires $O(N)$ storage and applying the forward operator scales requires $O(N^3 \log N)$ FLOPS, in contrast to dense matrix methods that require $O(N^4)$ storage and FLOPS. While the derivation and theoretical analysis of our algorithm is performed under the paraxial approximation, numerical evidence suggests the algorithm is accurate even when the paraxial approximation fails to hold. We retain the speed of earlier asymptotic algorithms while retaining enough accuracy to use iterative reconstruction methods.

4.2 Separable Approximation of $\hat{A}(\mathbf{k}_\parallel, k_0, z)$

With the exception of Section 4.4, we consider the single focal plane case. By translating the coordinate system we can take $z_F = 0$. We adopt the notation from the Chapter 3. We assume the source power spectrum $\left|\rho(k_0)\right|^2$ is supported on $[k_a, k_b]$.

First, we derive an approximation to the ISAM kernel (3.4) of the form

$$\hat{A}(\mathbf{k}_\parallel, k_0, z) \approx H(\mathbf{k}_\parallel, k_0) W(k_0, z) e^{\mathrm{i}\phi(\mathbf{k}_\parallel, k_0)z}, \tag{4.1}$$

and then further approximate this function as

$$\hat{A}(\mathbf{k}_\parallel, k_0, z) \approx H(\mathbf{k}_\parallel, k_0) w_1(k_0) \zeta_1(z) e^{\mathrm{i}\phi(\mathbf{k}_\parallel, k_0)z}. \tag{4.2}$$

Our approximated kernel (4.2) has the same functional form as the narrowband, paraxial kernel defined by Ralston *et al.* [99]; however, the details of the approximations differ.

We take $\rho(k_0) = \sqrt{k_0}$ for $k_0 \in [k_a, k_b]$ and zero otherwise. To begin, observe that the Gaussian function (3.1) can be written as a separable function in k_x and k_y. In particular, define

$$\begin{aligned}
\hat{g}_1(k_x, k_0) &\triangleq \sqrt{\frac{\rho(k_0)}{k_0 \mathrm{NA}}} \exp\left\{-\frac{|k_x|^2}{(k_0\mathrm{NA})^2}\right\} \\
\hat{g}(\mathbf{k}_\parallel, k_0) &= \frac{\rho(k_0)}{k_0 \mathrm{NA}} \exp\left\{-\frac{\left|\mathbf{k}_\parallel\right|^2}{(k_0\mathrm{NA})^2}\right\} \\
&= \hat{g}_1(k_x, k_0)\hat{g}_1(k_y, k_0).
\end{aligned}$$

For sufficiently small NA, the Gaussian functions decay fast enough in $\left|\mathbf{k}_\parallel\right|$ that the numerator of the integrand decays to zero well before the singularity at $\left|\mathbf{k}_\parallel\right|^2 = k_0^2$; thus, we can extend the limits of integration to ±∞. Under the paraxial approximation, we replace the k_z terms in the exponential of (3.4) by the quadratic approximation

$$k_z\left(\mathbf{k}_\parallel, k_0\right) \approx k_0 - \frac{k_x^2 + k_y^2}{2k_0} = k_0 - \frac{\left|\mathbf{k}_\parallel\right|^2}{2k_0},$$

while in the denominator of (3.4) we retain only the leading term. Thus, under the paraxial approximation, the kernel is given by

$$\begin{aligned}
\hat{A}(\mathbf{k}_\parallel, k_0, z) &= \int_{-\infty}^{\infty} \frac{g(\mathbf{k}'_\parallel, k_0) g(\mathbf{k}_\parallel - \mathbf{k}'_\parallel, k_0)}{k_0} \exp\left\{\mathrm{i}z\left(2k_0^2 - \frac{k_x'^2 + (k_x - k'_x)^2 + k_y'^2 + (k_y - k'_y)^2}{2k_0}\right)\right\} \mathrm{d}^2 k'_\parallel. \\
&= \frac{e^{\mathrm{i}2k_0 z}}{k_0} I_1(k_x, k_0, z) I_1(k_y, k_0, z),
\end{aligned}$$

where $I_1(k_x, k_0, z)$ is the one-dimensional integral

$$\begin{aligned}
I_1(k_x, k_0, z) &= \int_{-\infty}^{\infty} g_1(k'_x, k_0)\, g_1(k_x - k'_x, k_0) \exp\left\{-\mathrm{i}z \frac{k'^2_x + (k_x - k'_x)^2}{2k_0}\right\} \mathrm{d}k'_x \\
&= \frac{\rho(k_0)}{k_0 \mathrm{NA}} \int_{-\infty}^{\infty} \exp\left\{-\left(\frac{1}{k_0^2 \mathrm{NA}^2} + \frac{\mathrm{i}z}{2k_0}\right)(k'^2_x + (k_x - k'_x)^2)\right\} \mathrm{d}k'_x \\
&= \rho(k_0) \sqrt{\frac{\pi}{2}} \sqrt{\frac{1}{1 + \frac{\mathrm{i}}{2}\mathrm{NA}^2 k_0 z}} \exp\left\{-\frac{k_x^2}{2k_0^2 \mathrm{NA}^2} - \mathrm{i}\frac{k_x^2}{4k_0} z\right\},
\end{aligned}$$

and we have used the well-known Gaussian convolution identity $\int e^{-\tau(x-y)^2} e^{-\tau y^2} \mathrm{d}y = \sqrt{\frac{\pi}{2\tau}} e^{-\frac{\tau}{2}x^2}$.

Recall $\rho(k_0) = \sqrt{k_0}$ and the shorthand $|\mathbf{k}_\parallel| = \sqrt{k_x^2 + k_y^2}$. Continuing,

$$\begin{aligned}
\hat{A}(\mathbf{k}_\parallel, k_0, z) &= e^{\mathrm{i}2k_0 z} \frac{\rho(k_0)^2}{k_0} I_1(k_x, k_0, z) I_1(k_y, k_0, z) \\
&= e^{\mathrm{i}\frac{\pi}{2}2k_0 z} \left(\frac{1}{1 + \frac{\mathrm{i}}{2}\mathrm{NA}^2 k_0 z}\right) \exp\left\{-\frac{k_x^2 + k_y^2}{2k_0^2 \mathrm{NA}^2}\right\} \exp\left\{-\mathrm{i}\frac{k_x^2 + k_y^2}{4k_0} z\right\} \\
&= \frac{\pi}{2} \left(\frac{1}{1 + \frac{\mathrm{i}}{2}\mathrm{NA}^2 k_0 z}\right) \exp\left\{-\frac{|\mathbf{k}_\parallel|^2}{2k_0^2 \mathrm{NA}^2}\right\} \exp\left\{\mathrm{i}\left(2k_0 - \frac{|\mathbf{k}_\parallel|^2}{4k_0}\right) z\right\}.
\end{aligned}$$

We obtain the desired form (4.1) by defining the scalar

$$\gamma \triangleq \frac{\mathrm{NA}^2}{2}$$

and functions

$$\begin{aligned}
H(\mathbf{k}_\parallel, k_0) &\triangleq \frac{\pi}{2} \exp\left\{-\frac{|\mathbf{k}_\parallel|^2}{2k_0^2 \mathrm{NA}^2}\right\}, \\
W(k_0, z) &\triangleq \frac{1}{1 + \mathrm{i}\gamma k_0 z} \\
\phi(k_0, \mathbf{k}_\parallel) &\triangleq 2k_0 - \frac{|\mathbf{k}_\parallel|^2}{4k_0}.
\end{aligned} \tag{4.3}$$

Before continuing, note that $W(k_0, z)$ can be written in the magnitude and phase form

$$W(k_0, z) = \frac{1}{1 + \mathrm{i}\gamma k_0 z} = \frac{1}{\sqrt{1 + k_0^2 z^2 \gamma^2}} e^{-\mathrm{i}\,\arctan(\gamma k_0 z)}.$$

The phase term is known as the Gouy phase [98], and represents a phase shift of π as the focused Gaussian beam passes through the focal plane. This term is roughly $\pm\pi/2$ for any distance greater

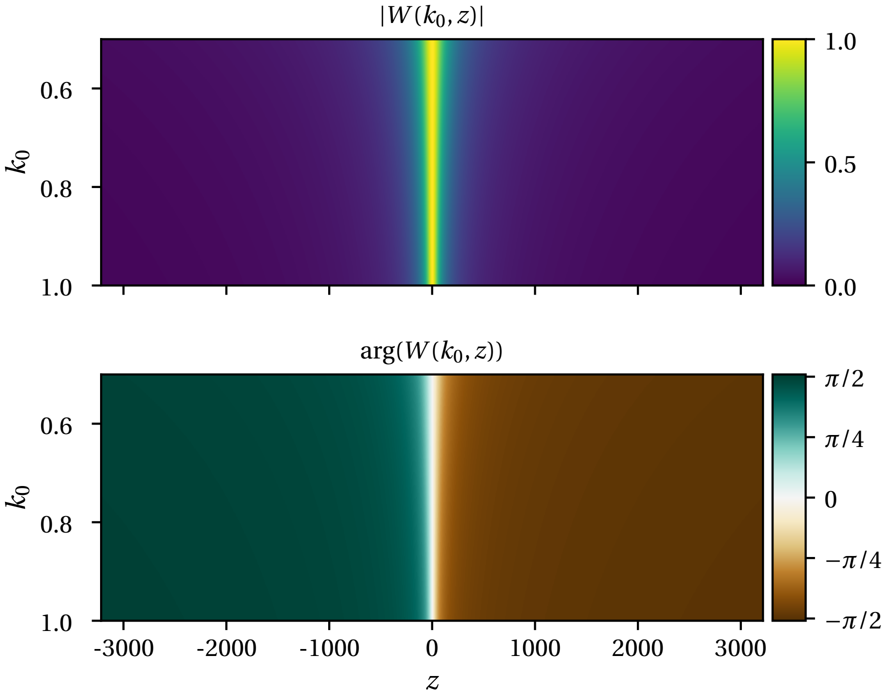


Figure 4.1: $W(k_0, z)$ for $\mathrm{NA} = 0.2$. The Gouy phase shift is evident in $\arg W(k_0, z)$.

than one Rayleigh range, $|z| > 1/(k_0\mathrm{NA}^2)$. The magnitude and phase of $W(k_0, z)$ for $\mathrm{NA} = 0.2$ is illustrated in Fig. 4.1.

4.2.1 Rank-one Approximation to $W(k_0, z)$

We have approximated $\hat{A}$ as the product of three terms: H, which is purely real; W, which has nonlinear phase but is smooth; and $e^{\mathrm{i}\phi(\mathbf{k}_\parallel, k_0)z}$, which is oscillatory but has linear phase in z. Next, we form a separable, or rank-one, approximation to $W(k_0, z)$ for $k_0 \in [k_a, k_b]$ and $z \in \mathbb{R}$. In particular, we want

$$W(k_0, z) \approx w_1(k_0)\zeta_1(z), \tag{4.4}$$

where $w_1 \in L^2([k_a, k_b])$ and $\zeta_1 \in L^2(\mathbb{R})$.

Define the operator $\mathcal{W} : L^2(\mathbb{R}) \to L^2([k_a, k_b])$ with kernel $W(k_0, z)$ by

$$(\mathcal{W}f)(k_0) = \int_{-\infty}^{\infty} W^*(k_0, z) f(z) \mathrm{d}z.$$

Observe that the kernel satisfies

$$\|W\|^2_{L^2([k_a,k_b]\times\mathbb{R})} = \int_{-\infty}^{\infty} \mathrm{d}z \int_{k_a}^{k_b} \mathrm{d}k_0\, |W(k_0,z)|^2 = \frac{\pi}{\gamma}\log\left(\frac{k_b}{k_a}\right) < \infty \tag{4.5}$$

implying $\mathcal{W}$ is of Hilbert-Schmidt type and thus has a singular value decomposition. The leading singular functions of $\mathcal{W}$ provide the best separable approximation to $W(k_0,z)$ in the L^2 sense. Unfortunately, these functions are not easy to compute; instead, we will find a function that is *almost* the leading left singular function. We make the *ansatz* that $W(k_0,z)$ can be well approximated by its value at some wavenumber $k_a \le \mu \le k_b$ and

$$\zeta_1(z) \triangleq \sqrt{\frac{\gamma\mu}{\pi}}\, W(\mu,z) = \sqrt{\frac{\gamma\mu}{\pi}}\frac{1}{1+\mathrm{i}\gamma\mu z}.$$

This choice satisfies $\int_{-\infty}^{\infty}|\zeta_1(z)|^2\,\mathrm{d}z = 1$. In the spirit of the singular value decomposition, define the function $w = \mathcal{W}\zeta_1$; explicitly,

$$\begin{aligned} w_1(k_0) = (\mathcal{W}\zeta_1)(k_0) &= \sqrt{\frac{\gamma\mu}{\pi}}\int_{-\infty}^{\infty}\big((1-\mathrm{i}\gamma k_0 z)(1+\mathrm{i}\gamma\mu z)\big)^{-1}\,\mathrm{d}z \\ &= \sqrt{\frac{\gamma\mu}{\pi}}\int_{-\infty}^{\infty}\left(1+\gamma^2 z^2 k_0\mu\right)^{-1}\mathrm{d}z \\ &= \sqrt{\frac{\pi\mu}{\gamma}}\frac{2}{\mu+k_0}. \end{aligned}$$

Note that $w_1(k_0)$ is real-valued. To avoid carrying around an additional scalar, we will not normalize w_1 to have unit norm. Our rank-one approximation to $W(k_0,z)$ is given by $w_1(k_0)\zeta_1(z)$.

The following proposition establishes that the best approximation to $W(k_0,z)$ occurs at the geometric mean of k_b and k_a.

Proposition 4.1. *The approximation error $\|W - w_1\zeta_1\|_{L^2[k_a,k_b]\times\mathbb{R})}$ is minimized at $\mu^\star = \sqrt{k_a k_b}$, and the corresponding relative approximation error is*

$$\frac{\|W - w_1\zeta_1\|^2_{L^2([k_a,k_b]\times\mathbb{R})}}{\|W\|^2_{L^2([k_a,k_b]\times\mathbb{R})}} = 1 - \frac{4\left(\frac{k_b}{k_a}-1\right)}{\left(1+\sqrt{\frac{k_b}{k_a}}\right)^2\log\left(\frac{k_b}{k_a}\right)}. \tag{4.6}$$

Proof. See Appendix B.2. □

The (relative) approximation error (4.6) depends only on the ratio of the maximum and minimum illumination wavenumbers, k_b and k_a. This quantity is plotted as a function of the ratio k_b/k_a in Fig. 4.2. Interestingly, the NA does not affect the relative approximation error. Note

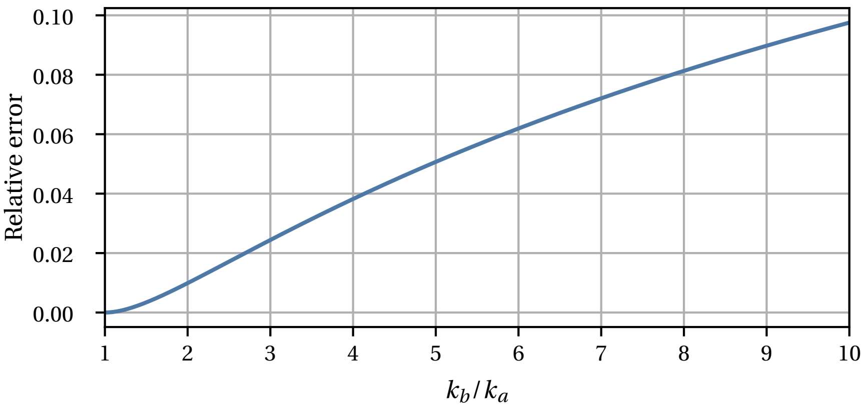


Figure 4.2: Relative approximation error (4.6) as a function of k_b/k_a.

that this result should be interpreted within the region of validity of the paraxial approximation, which *is* dependent on NA.

4.2.2 Beyond the Narrowband Regime: Rank r Approximation

For high-bandwidth systems the approximation can be improved by using a higher-order separable approximation, *i.e.*

$$W(k_0, z) = \sum_{j=1}^{r} w_j(k_0)\zeta_j(z). \tag{4.7}$$

We call (4.7) a *rank* r approximation, as we take the w_j and ζ_j to be the left and right singular functions of $\mathcal{W}$, respectively. An analytic expression for the higher-order singular functions may be possible using the variational characterization of the eigenfunctions of $\mathcal{W}\mathcal{W}^*$ and $\mathcal{W}^*\mathcal{W}$, but this is beyond the scope of this work.

The first three singular functions of $\mathcal{W}$ with $k_a = 1/2, k_b = 1$ are shown in Fig. 4.3. The singular functions were found numerically by forming a discretized approximation to $\mathcal{W}$, denoted $\mathbf{W} \in \mathbb{C}^{512\times4000}$, with $\mathbf{W}_{i,j} = W(k_a + i(k_a - k_b)/1024, -2000 + j)$ and performing the SVD.

The singular value ratios of $\mathbf{W}$ for various k_b are shown in Fig. 4.4. As expected, the singular values decay quickly for $k_b \approx k_a$, indicating that more terms of (4.7) are necessary in the high-bandwidth regime. Still, by including more terms, our approximation remains accurate even when the narrowband assumption fails.

Proposition 4.1 stands in contrast to the original paraxial, narrowband approximation of Ralston *et al.* [99]. They chose $\mu = (k_a + k_b)/2$, *i.e.* the arithmetic (rather than geometric) mean. Note, however, that their formulation of ISAM is scaled somewhat differently than ours, due to a different normalization for the Gaussian functions. Their equivalent of $W(k_0, z)$ carries an

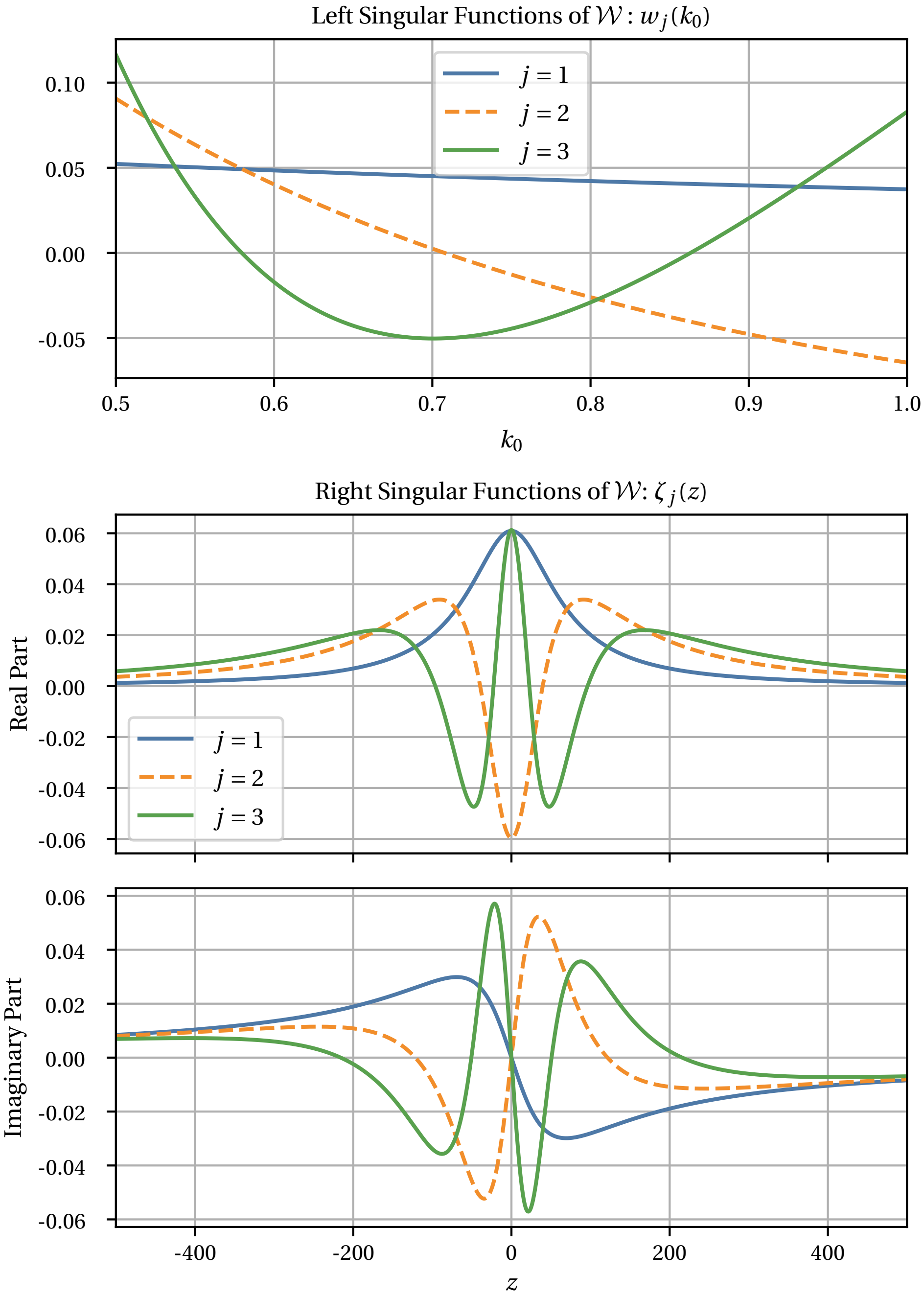


Figure 4.3: Left and right singular functions of **W**, the discretized form of $\mathcal{W}$, with NA = 0.2.

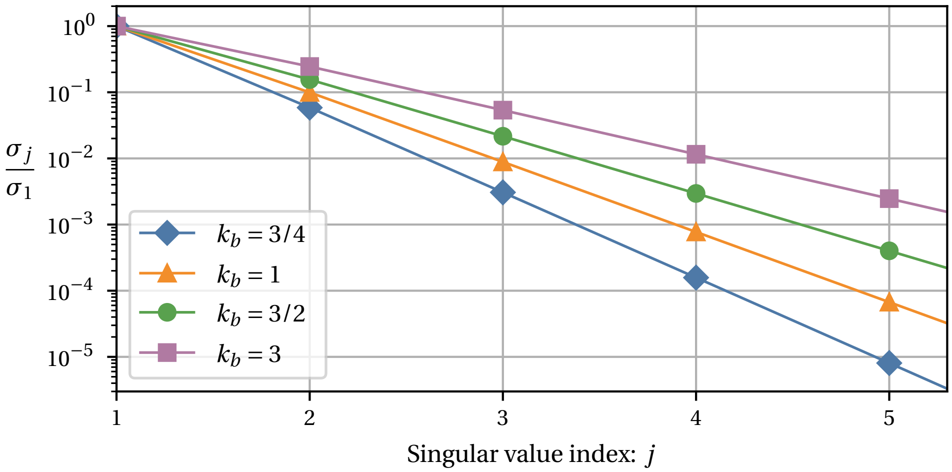


Figure 4.4: Ratio of the singular values of $\mathbf{W}$, the discretized form of $\mathcal{W}$ for increasing k_b. In each case, $k_a = 1/2$.

additional k_0^2 scaling compared to (4.3), and this influences the optimal approximation point. The optimum point is not easily determined using their scaling.

4.2.3 Beyond the Paraxial Approximation

Our analysis has been performed under the paraxial approximation. This approximation is necessary to have tractable integrals that lead to the factored form $\hat{A}(\mathbf{k}_\parallel, k_0, z)$ (4.1).

Recall from Section 3.4 that the when the asymptotic approximations are used, the ISAM kernel has a similar functional form to our (4.4). The primary difference is the behavior of the $\zeta(z)$ function and the phase function. In both the near and far from focus regimes, the term that has linear phase in z is given by $e^{\mathrm{i}\sqrt{4k_0^2 - |\mathbf{k}_\parallel|^2}z}$.

We postulate that (4.1), where $\phi(\mathbf{k}_\parallel, k_0) = \sqrt{4k_0^2 - \left|\mathbf{k}_\parallel\right|^2}$ and with W given by (4.3), is a good approximation to the ISAM kernel even when the paraxial approximation fails. We do not have a proof of this claim, although numerical evidence suggests it to be true.

To test this claim, we numerically calculated $\hat{A}(\mathbf{k}_\parallel = 0, k_0, z)$ for NA = 0.5 at $k_0 = 1/2$ and $k_0 = 1$ by evaluating the ISAM kernel integral (3.4). This is well beyond the paraxial approximation.

According to our claim, we have $\hat{A}(0, k_0, z) = W(k_0, z)e^{\mathrm{i}2k_0 z}$. To isolate $W(k_0, z)$, we multiplied by $e^{-\mathrm{i}2k_0 z}$. In Fig. 4.5 we plot the resulting function as well as our prediction, $W(k_0, z)$. There is excellent agreement between the prediction and the obtained function, leading us to believe the methods in this chapter apply beyond the paraxial approximation.

We postulate that (4.1) is a good approximation for all $\mathbf{k}_\parallel$, not just $\left|\mathbf{k}_\parallel\right| = 0$. To test this claim, we repeated the test using $\mathbf{k}_\parallel = (0.3, 0)$. According to the claim, we have $\hat{A}(\mathbf{k}_\parallel, k_0, z) = H(\mathbf{k}_\parallel, k_0) W(k_0, z) e^{\mathrm{i}\sqrt{4k_0^2 - |\mathbf{k}_\parallel|^2} z}$. We isolate $W(k_0, z)$ by dividing by $H(\mathbf{k}_\parallel, k_0)$ and multiplying by $e^{\mathrm{i}\sqrt{4k_0^2 - |\mathbf{k}_\parallel|^2} z}$. We plot the result in Fig. 4.6. Again, there is good agreement between the prediction and the obtained function.

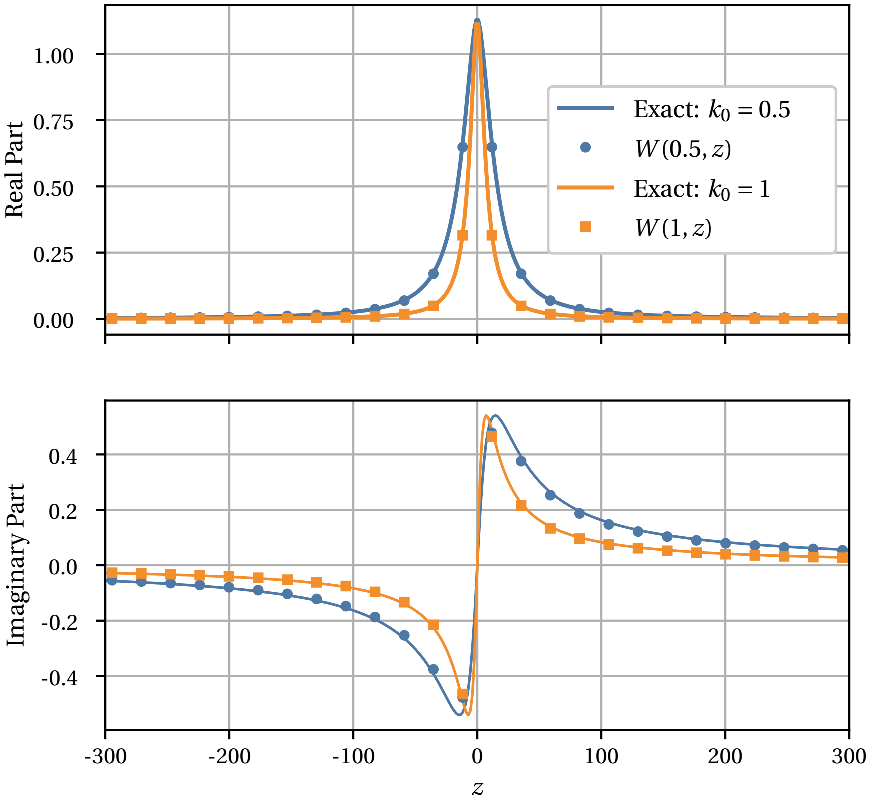


Figure 4.5: $W(k_0, z)$ and the "demodulated" function $\hat{A}(0, k_0, z) e^{-\mathrm{i}2k_0 z}$.

4.3 Singular Value Decomposition

4.3.1 Motivation and Related Work

We turn our attention to the study of the singular system of the ISAM operator, assuming the ISAM kernel is given by (4.2).

The singular value decomposition of a "imaging operator" provides tremendous insight into the capabilities of the imaging system. The singular system provides a characterization of the nullspace of the imaging operator— the components to which the imaging system is inherently

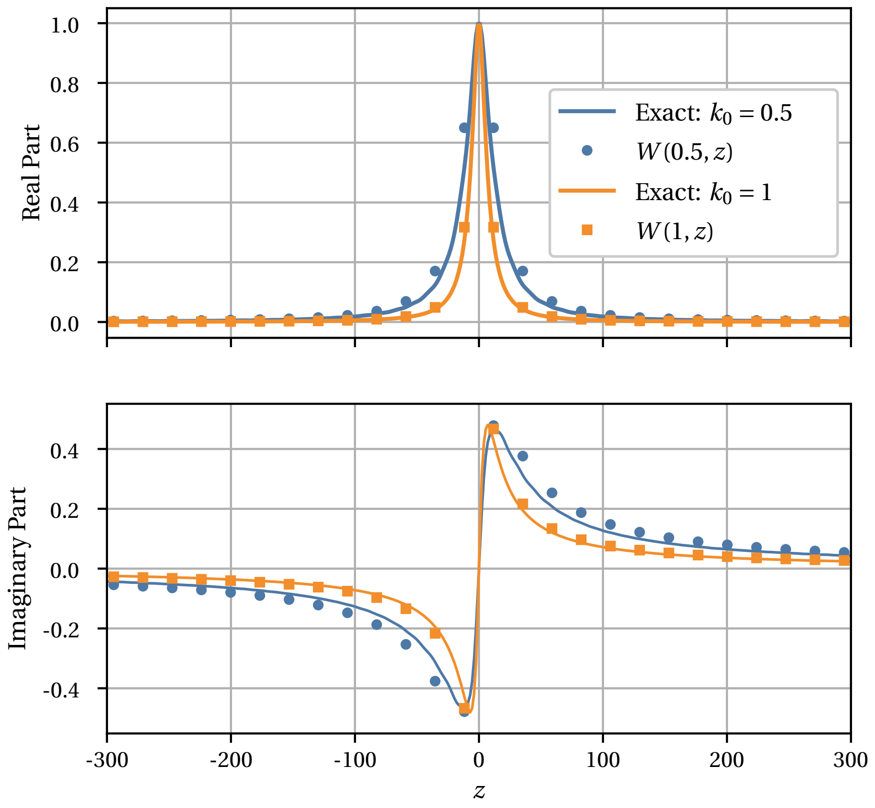


Figure 4.6: $W(k_0, z)$ and the "demodulated" function $\hat{A}(\mathbf{k}_\parallel, k_0, z)e^{-\mathrm{i}\sqrt{4k_0^2-|\mathbf{k}_\parallel|^2}z}/H(\mathbf{k}_\parallel, k_0)$, calculated using $\mathbf{k}_\parallel = (0.3, 0)$.

blind [117, 129–131]. Range conditions can be used to reduce noise and perturbations in measured data [132, 133]. Oscillation properties of the singular functions provide a way to define the "resolution" of an imaging system [120, 134–136].

Fix a value of $\mathbf{k}_\parallel$ and define the approximate ISAM operator $\tilde{A}_{\mathbf{k}_\parallel} : L^2(\mathbb{R}) \to L^2([k_a, k_b])$, which uses the rank-one approximation (4.4); we have

$$(\tilde{A}_{\mathbf{k}_\parallel} f)(k_0) = H(\mathbf{k}_\parallel, k_0)w_1(k_0)\int_{-\infty}^{\infty} e^{\mathrm{i}\phi(\mathbf{k}_\parallel, k_0)z}\zeta_1(z)f(\mathbf{k}_\parallel, z)\mathrm{d}z.$$

We will show that $\tilde{A}_{\mathbf{k}_\parallel}$ is compact and thus has a countable, discrete spectrum. The singular system of $\tilde{A}_{\mathbf{k}_\parallel}$ is the set of triples $\{\sigma_{\mathbf{k}_\parallel,j}, u_{\mathbf{k}_\parallel,j}, v_{\mathbf{k}_\parallel,j}\}_{j=1}^{\infty}$. The non-negative scalars $\sigma_{\mathbf{k}_\parallel,j}$ are the *singular values*. The singular values are ordered such that $\sigma_{\mathbf{k}_\parallel,1} \geq \sigma_{\mathbf{k}_\parallel,2}, \ldots \geq 0$. The functions $u_{\mathbf{k}_\parallel,j} \in L^2([k_a, k_b])$ are called the *left singular functions*, and $v_{\mathbf{k}_\parallel,j} \in L^2(\mathbb{R})$ are the *right singular* *functions.* The singular system satisfies the relations

$$(\tilde{\mathcal{A}}_{\mathbf{k}_\parallel} v_{\mathbf{k}_\parallel,j})(k_0) = \sigma_{\mathbf{k}_\parallel,j}\, u_{\mathbf{k}_\parallel,j}(k_0) \tag{4.8}$$

$$(\tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel} u_{\mathbf{k}_\parallel,j})(z) = \sigma_{\mathbf{k}_\parallel,j}\, v_{\mathbf{k}_\parallel,j}(z). \tag{4.9}$$

The left and right singular functions form orthonormal sets in $L^2([k_a,k_b])$ and $L^2(\mathbb{R})$, respectively. We include any singular functions with singular values equal to zero in our definition of the singular set; thus the left and right singular vectors form orthonormal bases for their respective spaces. Moreover, the set of $u_{\mathbf{k}_\parallel,j}$ with non-zero singular value forms an orthonormal basis for the range of $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}$. Similarly, the set of $v_{\mathbf{k}_\parallel,j}$ with non-zero singular value forms an orthonormal basis for the range of $\tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel}$ For this reason, the left and right singular functions are often called "data-space" and "object-space" singular functions, respectively.

Applying $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}$ to both sides of (4.9) yields

$$(\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}\tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel} u_{\mathbf{k}_\parallel,j})(k_0) = \sigma^2_{\mathbf{k}_\parallel,j}\, u_{\mathbf{k}_\parallel,j},$$

that is, the left singular vectors are eigenvectors $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}\tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel}$ with eigenvalue $\sigma^2_{\mathbf{k}_\parallel,j}$. We use this relation to study the singular system of $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}$. We find the left singular functions by solving this eigenvalue problem, and then obtain the right singular functions via (4.8).

In what follows, we we drop the $\mathbf{k}_\parallel$ from the subscript of the singular values and functions when it is clear that only a single $\mathbf{k}_\parallel$ is under consideration.

4.3.2 The Gram Operator

We begin by constructing the Gram operator, $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}\tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel}$. The kernel of this operator is

$$(\tilde{A}_{\mathbf{k}_\parallel}\tilde{A}^*_{\mathbf{k}_\parallel})(k_0,k_0') = H(\mathbf{k}_\parallel,k_0)H(\mathbf{k}_\parallel,k_0')\,w_1(k_0)\,w_1(k_0')\int_{-\infty}^{\infty}|\zeta_1(z)|^2\, e^{\mathrm{i}(\phi(\mathbf{k}_\parallel,k_0)-\phi(\mathbf{k}_\parallel,k_0'))z}\mathrm{d}z.$$

Concentrating on the integral;

$$\begin{aligned}\int_{-\infty}^{\infty} e^{\mathrm{i}(\phi(\mathbf{k}_\parallel,k_0)-\phi(\mathbf{k}_\parallel,k_0'))z}\,|\zeta_1(z)|^2\,\mathrm{d}z &= \int_{-\infty}^{\infty}\exp\left\{\mathrm{i}z\left(2k_0-\frac{|\mathbf{k}_\parallel|^2}{4k_0}-2k_0'+\frac{|\mathbf{k}_\parallel|^2}{4k_0'}\right)\right\}|\zeta_1(z)|^2\,\mathrm{d}z\\ &= \frac{\gamma\mu}{\pi}\int_{-\infty}^{\infty}\exp\left\{\mathrm{i}z\left(\frac{8k_0k_0'+|\mathbf{k}_\parallel|^2}{2k_0k_0'}(k_0-k_0')\right)\right\}(1+\gamma^2\mu^2z^2)^{-1}\mathrm{d}z\\ &= \exp\left\{\frac{-|k_0-k_0'|}{\gamma\mu}\left(2+\frac{|\mathbf{k}_\parallel|^2}{4k_0k_0'}\right)\right\},\end{aligned}$$

where we have used the fact that μ, γ, k_0, k_0' are each non-negative and the Fourier transform pair

$$\int_{-\infty}^{\infty} \frac{e^{-\mathrm{i}\omega z}}{1+z^2} \mathrm{d}z = \pi e^{-|\omega|}.$$

All together, we have

$$(\tilde{A}_{\mathbf{k}_\parallel} \tilde{A}^*_{\mathbf{k}_\parallel})(k_0, k_0') = H(\mathbf{k}_\parallel, k_0)\, w_1(k_0) \exp\left\{ \frac{-\left|k_0 - k_0'\right|}{\gamma\mu} \left(2 + \frac{\left|\mathbf{k}_\parallel\right|^2}{4k_0 k_0'}\right)\right\} H(\mathbf{k}_\parallel, k_0')\, w_1(k_0').$$

We introduce two additional operators to make the factorization clear. First, define the function

$$\tilde{G}_{\mathbf{k}_\parallel}(k_0, k_0') = \exp\left\{ \frac{-\left|k_0 - k_0'\right|}{\gamma\mu} \left(2 + \frac{\left|\mathbf{k}_\parallel\right|^2}{4k_0 k_0'}\right)\right\}, \tag{4.10}$$

along with the integral operator $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel} : L^2([k_a, k_b]) \to L^2([k_a, k_b])$,

$$(\tilde{\mathcal{G}}_{\mathbf{k}_\parallel} f)(k_0) = \int_{k_a}^{k_b} \tilde{G}_{\mathbf{k}_\parallel}(k_0, k_0') f(k_0') \mathrm{d}k_0'.$$

Define the function $d_{\mathbf{k}_\parallel}(k_0) \in L^2([k_a, k_b])$ by

$$d_{\mathbf{k}_\parallel}(k_0) \triangleq H(\mathbf{k}_\parallel, k_0)\, w_1(k_0),$$

and a "diagonal" operator $\mathcal{D}_{\mathbf{k}_\parallel} : L^2([k_a, k_b]) \to L^2([k_a, k_b])$ that scales the input by $d_{\mathbf{k}_\parallel}$; that is,

$$(\mathcal{D}_{\mathbf{k}_\parallel} f)(k_0) = H(\mathbf{k}_\parallel, k_0)\, w_1(k_0) f(k_0).$$

The function $d_{\mathbf{k}_\parallel}$ is continuous, real-valued, non-negative, bounded, and square-integrable; the operator $\mathcal{D}_{\mathbf{k}_\parallel}$ is bounded and self-adjoint. Using these operators, we have

$$\tilde{\mathcal{A}}_{\mathbf{k}_\parallel} \tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel} = \mathcal{D}_{\mathbf{k}_\parallel} \tilde{\mathcal{G}}_{\mathbf{k}_\parallel} \mathcal{D}_{\mathbf{k}_\parallel}.$$

Proposition 4.2. *The operators $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ and $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel} \tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel}$ are both self-adjoint, Hilbert-Schmidt and thus compact.*

Proof. See Appendix B.3. □

By the spectral theorem, the eigenfunctions of both $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel} \tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel}$ and $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ (including those corresponding to eigenvalues of zero) form an orthonormal basis for $L^2([k_a, k_b])$ [120, Theorem 3.19]. The eigenfunctions corresponding to the non-zero eigenvalues form a basis for the range of the operator.

Next, we show that the eigenfunctions of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ are the solution to a certain Sturm-Liouville differential equation. The eigenfunctions of $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}\tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel}$ are more challenging. Unfortunately, an eigendecomposition of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ does not directly lead to an eigendecomposition of $\mathcal{D}_{\mathbf{k}_\parallel}\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}\mathcal{D}_{\mathbf{k}_\parallel}$. Suppose we find the eigendecomposition for $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$, say, $\mathcal{U}\mathcal{T}\mathcal{U}^*$ where $\mathcal{U}$ is unitary and $\mathcal{T}$ is diagonal. Then $\mathcal{D}_{\mathbf{k}_\parallel}\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}\mathcal{D}_{\mathbf{k}_\parallel} = \mathcal{D}_{\mathbf{k}_\parallel}\mathcal{U}\mathcal{T}\mathcal{U}^*\mathcal{D}_{\mathbf{k}_\parallel}$, but the operator $\mathcal{D}_{\mathbf{k}_\parallel}\mathcal{U}^*$ is, in general, *not* unitary.

Still, an eigendecomposition of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ is useful, especially when coupled with the nice properties of our diagonal operators. Using Lemma B.4, if λ_n is the n-th eigenvalue of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ then the n-th eigenvalue of $\mathcal{D}_{\mathbf{k}_\parallel}\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}\mathcal{D}_{\mathbf{k}_\parallel}$, say ξ_n, is bounded between

$$\lambda_n \inf_{k_a<k_0<k_b} H(\mathbf{k}_\parallel, k_0)\, w_1(k_0) \le \xi_n \le \lambda_n \sup_{k_b<k_0<k_a} H(\mathbf{k}_\parallel, k_0)\, w_1(k_0). \tag{4.11}$$

Moreover, we can get a *generalized* eigendecomposition for $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}\tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel}$. We assume $\mathcal{D}^{-1}_{\mathbf{k}_\parallel}$ exists; otherwise, either 0 is the only point in the spectrum, or, owing to the monotonacity of the scaling function, we can restrict our attention from $L^2[k_a, k_b]$ to a region $L^2[k', k_b]$ where the inverse exists. Now, suppose φ is an eigenfunction of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ with eigenvalue λ; we have $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}\varphi = \lambda\varphi$. Let $\psi = \mathcal{D}^{-1}_{\mathbf{k}_\parallel}\varphi$. Then we have

$$\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}\tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel}\psi = \mathcal{D}_{\mathbf{k}_\parallel}\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}\mathcal{D}_{\mathbf{k}_\parallel}\mathcal{D}^{-1}_{\mathbf{k}_\parallel}\varphi = \lambda\mathcal{D}_{\mathbf{k}_\parallel}\varphi = \lambda\mathcal{D}_{\mathbf{k}_\parallel}\mathcal{D}_{\mathbf{k}_\parallel}\psi, \tag{4.12}$$

which is similar to the finite dimensional generalized eigenvalue problem $\mathbf{Ax} = \lambda\mathbf{Bx}$.

4.3.3 Conversion to a Sturm-Liouville Eigenvalue Problem

The following theorem identifies the eigenfunctions of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ with the solution of a certain Sturm-Liouville (SL) ordinary differential equation.

Theorem 4.1. *Let $f \in L^2[k_a, k_b]$ be an eigenfunction of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ with eigenvalue λ. Define*

$$p(k_0) \triangleq \frac{2\mathrm{NA}^2 \mu k_0^2}{|\mathbf{k}_\parallel|^2 + 8k_0^2}. \tag{4.13}$$

Then f is the unique solution to the regular Sturm-Liouville problem

$$-\frac{\mathrm{d}}{\mathrm{d}k_0}\left(p(k_0)\frac{\mathrm{d}f}{\mathrm{d}k_0}\right)+\frac{1}{p(k_0)}f=\frac{2}{\lambda}f, \tag{4.14}$$

$$f(k_a)-p(k_a)f'(k_a)=0, \tag{4.15}$$

$$f(k_b)+p(k_b)f'(k_b)=0 \tag{4.16}$$

over the interval $[k_a, k_b]$.

Proof. See Appendix B.3.1. □

The eigenvalue problem (4.14), with boundary conditions (4.15) and (4.16), is the simplest type of a SL problem. The function $p(k_0)$ is in $C^1([k_a, k_b])$ and does not change sign, the problem is over a finite interval $[k_a, k_b]$, and the boundary conditions are separated. Such a problem is called a *regular, self-adjoint Sturm-Liouville problem.*

Consider the differential operator $\mathcal{L} : C^2[k_a, k_b] \to C^2[k_a, k_b]$

$$\mathcal{L}f \triangleq -\frac{\mathrm{d}}{\mathrm{d}k_0}\left(p(k_0)\frac{\mathrm{d}f}{\mathrm{d}k_0}\right)+\frac{1}{p(k_0)}f.$$

Theorem 4.1 states that the eigenfunctions of $\mathcal{L}$ are also eigenfunctions as $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$, and the eigenvalues of $\mathcal{L}$ are inversely proportional to those of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$. Indeed, if the domain of $\mathcal{L}$ is restricted to functions that satisfy the boundary conditions (4.15) and (4.16), we have that $\mathcal{L}$ and $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ are the inverse of one another. Put another way, $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ is the Green's function for the SL problem (4.14) to (4.16).

The approach we have taken— obtaining the singular system of a linear operator by transitioning to the analysis of a SL eigenvalue problem— brings to mind the seminal work of Slepian, Landau and Pollak [137–141]. They sought to find eigenfunctions of a certain integral operator, related to space-and-frequency limited Fourier measurements. They found a certain well-studied SL differential operator that commutes with their integral operator, and thus obtained the eigensystem by analyzing the resulting differential equation.

The following result is a direct application of standard results on regular, self-adjoint SL problems; *e.g.* [142, Theorem 4.3.1]. We say two sequences $\{f_n\}$ and $\{g_n\}$ are *asymptotically equivalent as $n \to \infty$* if $\lim_{n\to\infty} \frac{f_n}{g_n} = 1$.

Corollary 4.2. *Let f be an eigenfunction of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ with eigenvalue λ.*

(P1) The eigenvalues of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ are countable, real, and can be ordered $\lambda_1 > \lambda_2 > \ldots > 0$.

(P2) The (normalized) eigenfunctions of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ form an orthonormal basis for $L^2[k_a, k_b]$.

(P3) The eigenvalues λ_n are asymptotically equivalent to n^{-2} as $n \to \infty$.

(P4) The n-th eigenfunction has exactly $n-1$ zeros in $[k_a, k_b]$.

Properties $(1-3)$ are expected and can be deduced from properties of compact, self-adjoint integral operators [143]. Property (4), however, is new, and has implications to the resolution of the imaging system.

In general, the solutions to the SL problem described in Theorem 4.1 cannot be given in closed form. Fortunately, the eigenvalue problem can be easily solved numerically using established methods, such as `SLEIGN2` [144].

4.3.4 How Does the Leading Eigenvalue Vary With $\mathbf{k}_\|$?

We are interested in the behavior of the eigenvalues of $\tilde{\mathcal{G}}_{\mathbf{k}_\|}$ as a function of $\mathbf{k}_\|$ and the imaging parameters k_b, k_a, and NA. Brown *et al.* developed bounds for eigenvalues of regular, self-adjoint Sturm-Liouville problems based on a combination of Soblev inequality and the Prüfer transformation. Translating their results to our problem, we have [145, Theorem 2.3 and Remark 2.3]

$$(n-1)^2 \le \frac{1}{4}\left(\int_{k_a}^{k_b} \frac{1}{p(k_0)} \mathrm{d}k_0\right)\left(\int_{k_a}^{k_b} \max\left\{\frac{2}{\lambda_n} - \frac{1}{p(k_0)}, 0\right\} \mathrm{d}k_0\right) \qquad \text{for } n \ge 2,$$

and for $n = 1$,

$$1 \le \frac{1}{4}\left(\int_{k_a}^{k_b} \frac{1}{p(k_0)} \mathrm{d}k_0\right)\left(\int_{k_a}^{k_b} \max\left\{\frac{2}{\lambda_1} - \frac{1}{p(k_0)}, 0\right\} \mathrm{d}k_0\right). \tag{4.17}$$

The inequality (4.17) gives insight into how the leading eigenvalue, behaves as function of the system parameters. The right side is non-zero if $2p(k_0) > \lambda_{\mathbf{k}_\|,1}$ for at least one $k \in [k_a, k_b]$. As $p(k_0)$ is monotone increasing, the right side is non-zero if and only if

$$\lambda_1 \le 2p(k_b) = \frac{2\mathrm{NA}^2(k_a + k_b)k_b^2}{\left|\mathbf{k}_\|\right|^2 + 8k_b^2}.$$

Thus the leading eigenvalue is a decreasing function of $\left|\mathbf{k}_\|\right|^2$ and is $O\left(\left|\mathbf{k}_\|\right|^{-2}\right)$. Interestingly, changing NA affects all leading eigenvalues uniformly; there is no coupling between $\left|\mathbf{k}_\|\right|^2$ and NA. We must be careful to remember that this behavior describes the eigenvalues of $\tilde{\mathcal{G}}_{\mathbf{k}_\|}$, not $\tilde{\mathcal{A}}_{\mathbf{k}_\|}$; indeed, we know from (4.11) that the leading eigenvalue of $\tilde{\mathcal{A}}_{\mathbf{k}_\|}$ is $O\left(e^{-\left|\mathbf{k}_\|\right|^2/\mathrm{NA}^2}\right)$.

Next, we describe two special cases in which the eigenfunctions can be found in closed form. First, we consider the special case of $\mathbf{k}_\| = 0$; afterwards, we return to the narrowband approximation.

4.3.5 Special Case: $\tilde{\mathcal{G}}_0$

The eigenvalue problem is considerably simplified when $|\mathbf{k}_{\|}| = 0$. In this case, the coefficient function $p(k_0)$ reduces to

$$p(k_0) = \frac{\mathrm{NA}^2\mu}{4};$$

i.e., a constant function of k_0. The SL problem becomes (4.14) becomes

$$\begin{aligned} -\frac{\mathrm{NA}^2\mu}{4} f'' + \left(\frac{4}{\mathrm{NA}^2\mu} - \frac{2}{\lambda}\right) f &= 0, \\ f(k_a) - \frac{\mathrm{NA}^2}{4} f'(k_a) &= 0, \\ f(k_b) + \frac{\mathrm{NA}^2}{4} f'(k_b) &= 0. \end{aligned} \tag{4.18}$$

Theorem 4.3. *The n-th (unnormalized) eigenfunction of $\tilde{\mathcal{G}}_0$ is*

$$f_n(k_0) = \begin{cases} \cos\left(\tau_n\left(k_0 - k_{\mathrm{mid}}\right)\right) \ \textit{where}\ \tau_n = -\dfrac{4}{\mu \mathrm{NA}^2}\cot\left(\tau_n \dfrac{k_a - k_b}{2}\right) > 0 & n \textit{ odd;} \\ \sin\left(\tau_n\left(k_0 - k_{\mathrm{mid}}\right)\right) \ \textit{where}\ \tau_n = \dfrac{4}{\mu \mathrm{NA}^2}\tan\left(\tau_n \dfrac{k_a - k_b}{2}\right) > 0 & n \textit{ even}, \end{cases}$$

where $\tau_n > \tau_i$ for $i = 1, \ldots, n-1$ and $k_{\mathrm{mid}} \triangleq (k_b + k_a)/2$. The n-th eigenvalue is

$$\lambda_n = \frac{8\mu \mathrm{NA}^2}{16 + \tau_n^2 \mu^2 \mathrm{NA}^4}. \tag{4.19}$$

Proof. See Appendix B.4. □

As the eigenvalues are determined by the solution of a transcendental equation, we cannot write them in closed form. However, we can bracket each solution of the transcendental equation into intervals of width $\pi/(k_b - k_a)$, thus providing upper and lower bounds for each eigenvalue.

Corollary 4.4. *The eigenvalues of $\tilde{\mathcal{G}}_0$ satisfy*

$$\frac{16(k_b - k_a)^2(k_b + k_a)\mathrm{NA}^2}{(n+1)^2\pi^2\mathrm{NA}^4(k_a + k_b)^2 + 64(k_b - k_a)^2} \le \lambda_n \le \frac{64(k_b - k_a)^2(k_b + k_a)\mathrm{NA}^2}{n^2\pi^2\mathrm{NA}^4(k_a + k_b)^2 + 64(k_b - k_a)^2}.$$

Proof. The result follows by combining the interval bounds of Lemma B.3 with (4.19). □

4.3.6 Special Case: Approximate Eigenfunctions in the Narrowband Regime

In general, the solutions of SL problem (4.14) cannot be expressed in closed form. However, the solutions to a closely related problem can be written in closed form. We simplify the problem by linearizing the coefficient function $p(k_0)$ about the point $k = \mu$ and solve the resulting SL eigenvalue problem. Let $\tilde{p}(k_0)$ denote the linearization of $p(k_0)$ about $k = \mu$. We have

$$p(k_0) = \tilde{p}(k_0) + O\left((k_0 - \mu)^2\right),$$

thus $\tilde{p}(k_0)$ is a good approximation to when $\left|k_0 - \mu\right|$ is small; this is the narrowband regime. The linearized function $\tilde{p}(k_0)$ is given by

$$\begin{aligned}\tilde{p}(k_0) &\triangleq \frac{2\mu^3\mathrm{NA}^2(8\mu^2 - \left|\mathbf{k}_\parallel\right|^2)}{(8\mu^2 + \left|\mathbf{k}_\parallel\right|^2)^2} + \frac{4\mu^2\mathrm{NA}^2\left|\mathbf{k}_\parallel\right|^2}{(8\mu^2 + \left|\mathbf{k}_\parallel\right|^2)^2}k_0 \\ &= \beta\left(\alpha + 2\left|\mathbf{k}_\parallel\right|^2 k_0\right),\end{aligned}$$

where we introduced scalars

$$\begin{aligned}\alpha &\triangleq 8\mu^3 - \mu\left|\mathbf{k}_\parallel\right|^2, \\ \beta &\triangleq \frac{2\mu^2\mathrm{NA}^2}{\left(8\mu^2 + \left|\mathbf{k}_\parallel\right|^2\right)^2}.\end{aligned}$$

Spectral properties of perturbed Sturm-Liouville problems have been studied. In the case of regular, self-adjoint problems with separated boundary conditions, such as the problem considered in Theorem 4.5, the n-th eigenvalue is a continuous function of $\tilde{p}$ and $1/\tilde{p}$. Under minor additional conditions the eigenvalue is also a continuous function of the boundary conditions [142, Section 4.4] [146]. These results imply the linearized problem is a good surrogate for the true eigenvalue problem.

Theorem 4.5 (Solutions to Linearized Sturm-Liouville Differential Equation)**.** *Let $\tilde{f}, \tilde{\lambda}$ be such that*

$$\begin{aligned}-\frac{\mathrm{d}}{\mathrm{d}k_0}\left(\tilde{p}(k_0)\frac{\mathrm{d}\tilde{f}}{\mathrm{d}k_0}\right) + \frac{1}{\tilde{p}(k_0)}\tilde{f} &= \frac{2}{\tilde{\lambda}}\tilde{f} \qquad (4.20)\\ \tilde{f}(k_a) - \tilde{p}(k_a)\tilde{f}'(k_a) &= 0, \\ \tilde{f}(k_b) + \tilde{p}(k_b)\tilde{f}'(k_b) &= 0.\end{aligned}$$

There exist unique scalars c_1, c_2 such that

$$\tilde{f}(k_0) = c_1 J\left(\frac{1}{\beta\left|\mathbf{k}_\parallel\right|^2}, \frac{1}{\left|\mathbf{k}_\parallel\right|^2}\sqrt{\frac{2(\alpha+2\left|\mathbf{k}_\parallel\right|^2 k_0)}{\beta\tilde{\lambda}}}\right) + c_2 Y\left(\frac{1}{\beta\left|\mathbf{k}_\parallel\right|^2}, \frac{1}{\left|\mathbf{k}_\parallel\right|^2}\sqrt{\frac{2(\alpha+2\left|\mathbf{k}_\parallel\right|^2 k_0)}{\beta\tilde{\lambda}}}\right), \tag{4.21}$$

where $J(n,x)$ is the n-th order Bessel function of the first kind, and $Y(n,x)$ is the n-th order Bessel function of the second kind.

Proof. See Appendix B.5. □

Unlike the case of $\left|\mathbf{k}_\parallel\right| = 0$, we cannot say that $c_1 \neq 0$ implies $c_2 = 0$ and vice versa. By standard properties of regular, self-adjoint Sturm-Liouville problems, there are only countably many choice of $\tilde{\lambda}$ for which the function (4.21) can satisfy the boundary conditions [142, Theorem 4.3.1]. Fortunately, $\tilde{\lambda}$ can be easily determined by numerical methods by way of the Prüfer transformation [147, Section 4.3.5].

Still, we can get some insight from the form of the solution (4.21). The eigenvalues must decay to zero as n tends to infinity. For sufficiently small $\tilde{\lambda}$, we can use the large argument asymptotic expansion of Bessel functions, namely [148, Equations 10.17.3, 10.17.4]. We combine the leading order of the large-argument expansion with the quadratic approximation

$$\sqrt{\alpha+2\left|\mathbf{k}_\parallel\right|^2 k_0} = \sqrt{8\mu^3+\mu\left|\mathbf{k}_\parallel\right|^2} + \frac{\left|\mathbf{k}_\parallel\right|^2}{\sqrt{8\mu^3+\mu\left|\mathbf{k}_\parallel\right|^2}}(k_0-\mu) + O\big((k_0-\mu)^2\big),$$

which is accurate in the narrowband regime. We obtain

$$J\left(\frac{1}{\beta\left|\mathbf{k}_\parallel\right|^2}, \frac{1}{\left|\mathbf{k}_\parallel\right|^2}\sqrt{\frac{2(\alpha+2\left|\mathbf{k}_\parallel\right|^2 k_0)}{\beta\tilde{\lambda}}}\right) \sim \left(\frac{2\beta\tilde{\lambda}}{\alpha+2k_0\left|\mathbf{k}_\parallel\right|^2}\right)^{\frac{1}{4}} \cos\left(c_2 + \frac{(k_0-\mu)}{\sqrt{\tilde{\lambda}\beta\left(8\mu^3+\mu\left|\mathbf{k}_\parallel\right|^2\right)}}\right),$$

$$Y\left(\frac{1}{\beta\left|\mathbf{k}_\parallel\right|^2}, \frac{1}{\left|\mathbf{k}_\parallel\right|^2}\sqrt{\frac{2(\alpha+2\left|\mathbf{k}_\parallel\right|^2 k_0)}{\beta\tilde{\lambda}}}\right) \sim \left(\frac{2\beta\tilde{\lambda}}{\alpha+2k_0\left|\mathbf{k}_\parallel\right|^2}\right)^{\frac{1}{4}} \sin\left(c_2 + \frac{(k_0-\mu)}{\sqrt{\tilde{\lambda}\beta\left(8\mu^3+\mu\left|\mathbf{k}_\parallel\right|^2\right)}}\right)$$

as $\tilde{\lambda}$ tends to zero. Here, c_1 and c_2 are constants that do not depend on $\tilde{\lambda}$. Note $(\alpha+2k_0\left|\mathbf{k}_\parallel\right|^2)^{-1/4}$ is slowly varying, especially compared to the oscillatory trigonometric terms, and can be neglected whenever $\left|\mathbf{k}_\parallel\right|$ is small. All together, these approximations imply that the high-order eigenfunctions of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ are well-approximated by trigonometric functions.

Table 4.1: Parameters for eigenvalue examples.

k_a	$0.5\ \mathrm{rad}\cdot\mu\mathrm{m}^{-1}$	L_x	$402.1\ \mu\mathrm{m}$	L_z	$3217\ \mu\mathrm{m}$
k_b	$1\ \mathrm{rad}\cdot\mu\mathrm{m}^{-1}$	N_x	128	N_z	4096
N_k	384	Δ_x	$\pi\ \mu\mathrm{m}$	Δ_z	$0.25\pi\ \mu\mathrm{m}$
NA	0.2				

4.3.7 Examples

We demonstrate the results of this section using numerical examples. We want to compare the eigensystem found by solving the Sturm-Liouville problem against the eigensystem of the integral operators $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ and left singular functions of $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}$.

We compute the solution to the Sturm-Liouville using SLEIGN2 [144]. We calculated the eigensystem and singular system of discretized approximations to $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ and $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}$. We formed a matrix containing samples of the corresponding kernels and then computed the SVD of this matrix. The discrete version of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ was formed by sampling (4.10). Importantly, we formed the discretized approximation to $\mathcal{A}_{\mathbf{k}_\parallel}$ using the exact ISAM kernel (3.3); that is, $\mathcal{A}_{\mathbf{k}_\parallel}$ was not formed using the paraxial approximation. The discretization parameters are listed in Table 4.1.

As $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}\tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel} = \mathcal{D}_{\mathbf{k}_\parallel}\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}\mathcal{D}_{\mathbf{k}_\parallel}$, we should have $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel} = \mathcal{D}^{-1}_{\mathbf{k}_\parallel}\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}\tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel}\mathcal{D}^{-1}_{\mathbf{k}_\parallel}$ if $\mathcal{D}_{\mathbf{k}_\parallel}$ is invertible. However, as $\mathcal{A}_{\mathbf{k}_\parallel}$ is not formed using the paraxial approximation, we do not have $\mathcal{A}_{\mathbf{k}_\parallel}\mathcal{A}^*_{\mathbf{k}_\parallel} = \mathcal{D}_{\mathbf{k}_\parallel}\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}\mathcal{D}_{\mathbf{k}_\parallel}$. We use the difference between $\mathcal{D}^{-1}_{\mathbf{k}_\parallel}\mathcal{A}_{\mathbf{k}_\parallel}\mathcal{A}^*_{\mathbf{k}_\parallel}\mathcal{D}^{-1}_{\mathbf{k}_\parallel}$ and $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ to evaluate robustness to the paraxial approximation.

First, we consider the case $\mathbf{k}_\parallel = 0$. We compare the eigenvalues predicted by solving the Sturm-Liouville problem (4.18) with the eigenvalues of $\tilde{\mathcal{G}}_0$ and those of $\mathcal{D}_0^{-1}\mathcal{A}_0\mathcal{A}_0^*\mathcal{D}_0^{-1}$. The eigenvalues are plotted in Fig. 4.7. The eigenvalues of $\tilde{\mathcal{G}}_0$ coincide with those found using SLEIGN2. The first five eigenvalues of $\tilde{\mathcal{G}}_0$ are lower than the corresponding eigenvalues of $\mathcal{D}_0^{-1}\mathcal{A}_0\mathcal{A}_0^*\mathcal{D}_0^{-1}$.

The eigenfunctions of $\tilde{\mathcal{G}}_0$ and $\mathcal{A}_0\mathcal{A}_0^*$ are shown in Fig. 4.8. As expected, the eigenfunctions of $\tilde{\mathcal{G}}_0$ are sinuosidal. The generalized eigenfunction relationship (4.12) is demonstrated in Fig. 4.9. Clearly, $\mathcal{D}_0\mathcal{D}_0 u_{0,15}(k_0) \approx \mathcal{A}_0\mathcal{A}_0^* u_{0,15}(k_0)$.

Next, we consider $\mathbf{k}_\parallel = (0.15, 0)$. In this case, we also computed the eigensystem using the linearized approximation discussed in Section 4.3.6. The eigenvalues of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ and $\mathcal{D}^{-1}_{\mathbf{k}_\parallel}\mathcal{A}_{\mathbf{k}_\parallel}\mathcal{A}^*_{\mathbf{k}_\parallel}\mathcal{D}^{-1}_{\mathbf{k}_\parallel}$, along with the eigenvalues found using the Sturm-Liouville solution, are shown in Fig. 4.10. The eigenvalues found by solving linearized Sturm-Liouville problem are in agreement with the non-linearized form and with the eigenvalues of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$. The eigenfunctions are shown in Fig. 4.11; the linearized and non-linearized versions are in good agreement. Although the first eigenfunction of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ does not look particularly sinusoidal, the 6th and 16th do. This is in agreement with the previous section. The eigenfunctions of $\mathcal{A}_{\mathbf{k}_\parallel}\mathcal{A}^*_{\mathbf{k}_\parallel}$ are markedly different. However, Fig. 4.12 shows

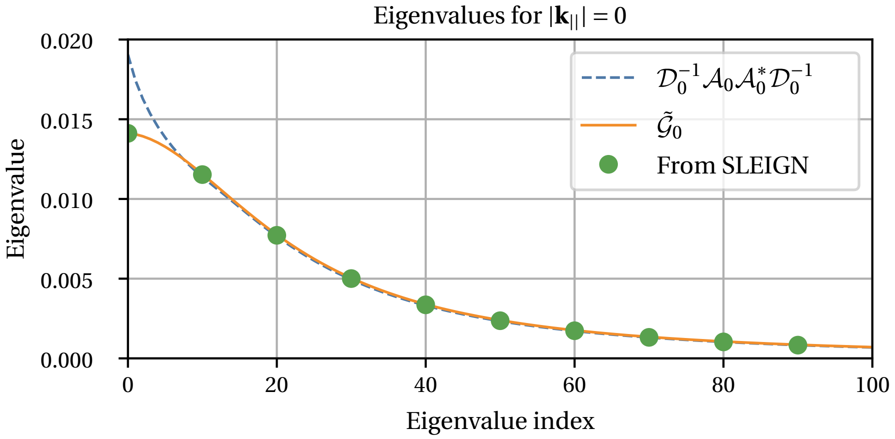


Figure 4.7: Eigenfunctions of $\tilde{\mathcal{G}}_0$ computed according to Theorem 4.3 and discretized approximations to $\tilde{\mathcal{G}}_0$ and $\mathcal{A}_0\mathcal{A}_0^*$. Here, $\mathbf{k}_{\|} = (0.15, 0)$.

the generalized eigenfunction relationship (4.12) still holds.

4.4 Numerical Methods

We now turn our attention to the practical implications of our rank-r approximation to the weighting function (4.7). We begin discussing the application of the forward ISAM operator, $\mathcal{A}$, and its adjoint, $\mathcal{A}^*$. We then apply our algorithm to the task of multi-focal ISAM reconstruction.

4.4.1 Forward and Adjoint Operators

Let $f \in L^2(\mathbb{R})$ be a compactly supported, non-dispersive object.

We define an approximate ISAM operator $\tilde{\mathcal{A}} : L^2(\mathbb{R}^3) \times L^2(\mathbb{R}^3)$ that uses our rank-r approximation to $W(k_0, z)$;

$$(\tilde{\mathcal{A}} f)(\mathbf{k}_{\|}, k_0) = \sum_{j=1}^{r} H(\mathbf{k}_{\|}, k_0)\, w_j(k_0) \int_{-\infty}^{\infty} \zeta_j(z) f(\mathbf{k}_{\|}, z) e^{\mathrm{i}\phi(\mathbf{k}_{\|}, k_0)z} \mathrm{d}z. \tag{4.22}$$

The adjoint of this operator is

$$(\tilde{\mathcal{A}}^* \varphi)(\mathbf{k}_{\|}, z) = \sum_{j=1}^{r} \zeta_j^*(z) \int_{k_a}^{k_b} H(\mathbf{k}_{\|}, k_0)\, w_j(k_0) \varphi(\mathbf{k}_{\|}, k_0) e^{-\mathrm{i}\phi(\mathbf{k}_{\|}, k_0)z} \mathrm{d}z. \tag{4.23}$$

Evaluation of (4.22) and (4.23) requires only pointwise multiplication, vector addition, and the

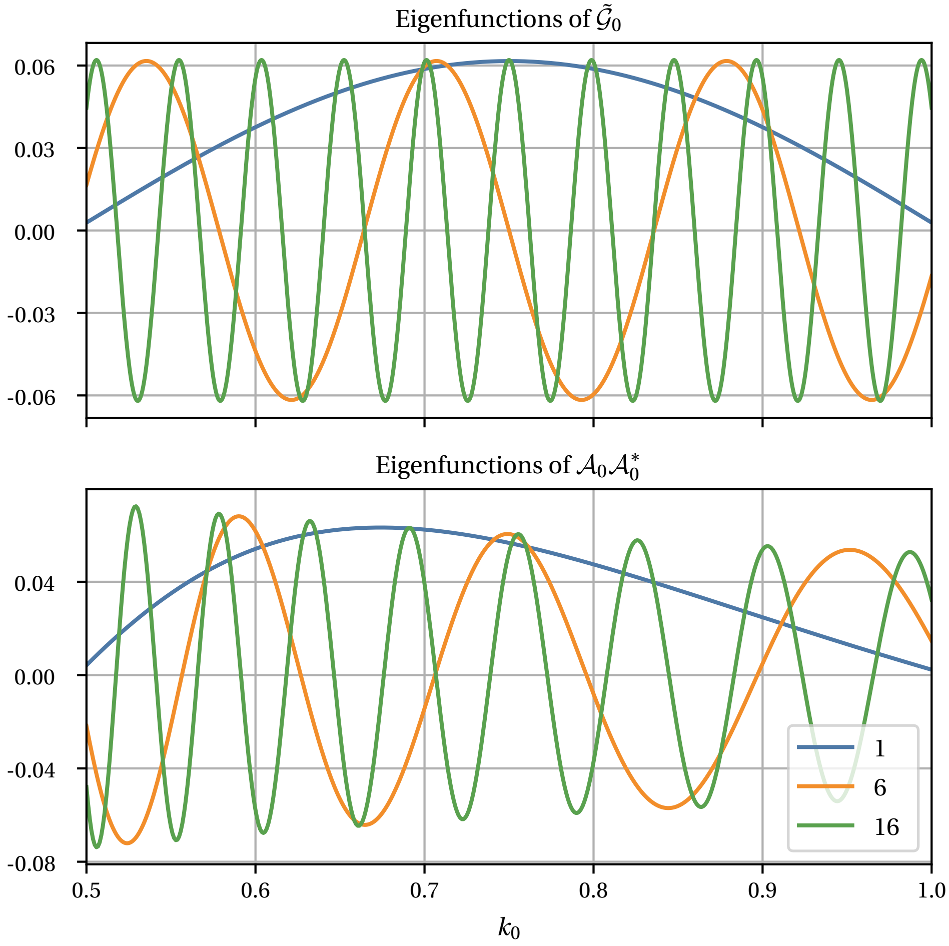


Figure 4.8: Selected eigenfunctions of $\tilde{\mathcal{G}}_0$ and $\mathcal{A}_0\mathcal{A}_0^*$.

computation of the Fourier transform at unevenly spaced locations; in particular, we need the Fourier transform variable at locations $k_z = \phi(\mathbf{k}_\parallel, k_0)$. This final task can be accomplished using Non-Uniformly spaced FFT (NUFFT) or Unevenly Spaced FFT (USFFT) algorithms [149–151], but for simplicity we apply a zero-padded FFT and use linear interpolation to determine the Fourier transform at the desired locations.

The proposed algorithm for the forward operator is given in Algorithm 4 and the proposed algorithm for the adjoint is listed in Algorithm 5.

Note that care must be taken with the resampling steps of Algorithms 4 and 5. The resampling step used in the adjoint operation must be adjoint to the interpolation step used in the forward operator. The resampling operation in the application of Algorithm 5 should be *anterpolation*, rather than interpolation.

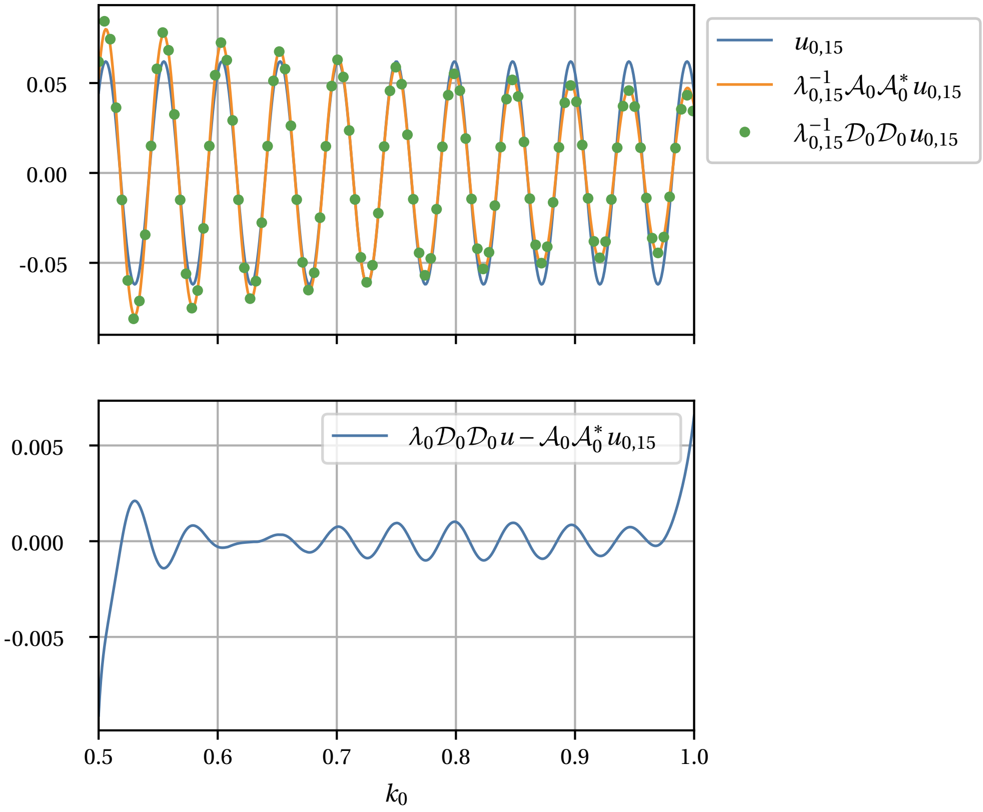


Figure 4.9: Demonstrating the generalized eigenfunction relationship (4.12).

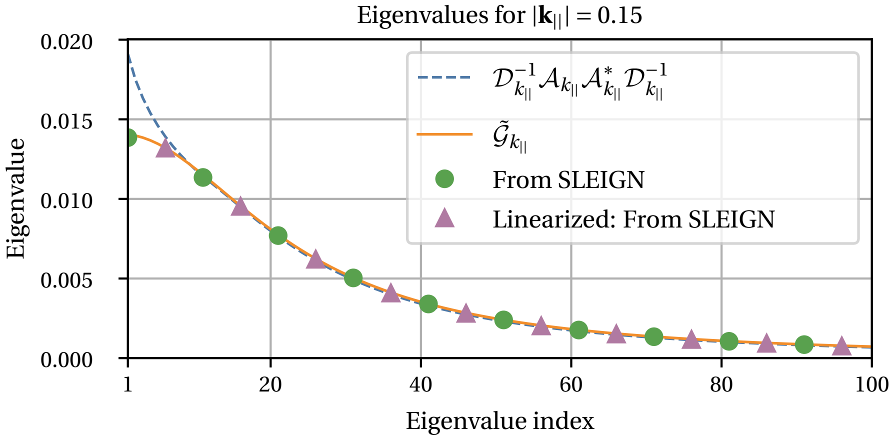


Figure 4.10: Eigenfunctions of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ computed according to (4.14), (4.20), and discretized approximations to $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ and $\mathcal{A}_{\mathbf{k}_\parallel}\mathcal{A}_{\mathbf{k}_\parallel}^*$. Here, $\mathbf{k}_\parallel = (0.15, 0)$.

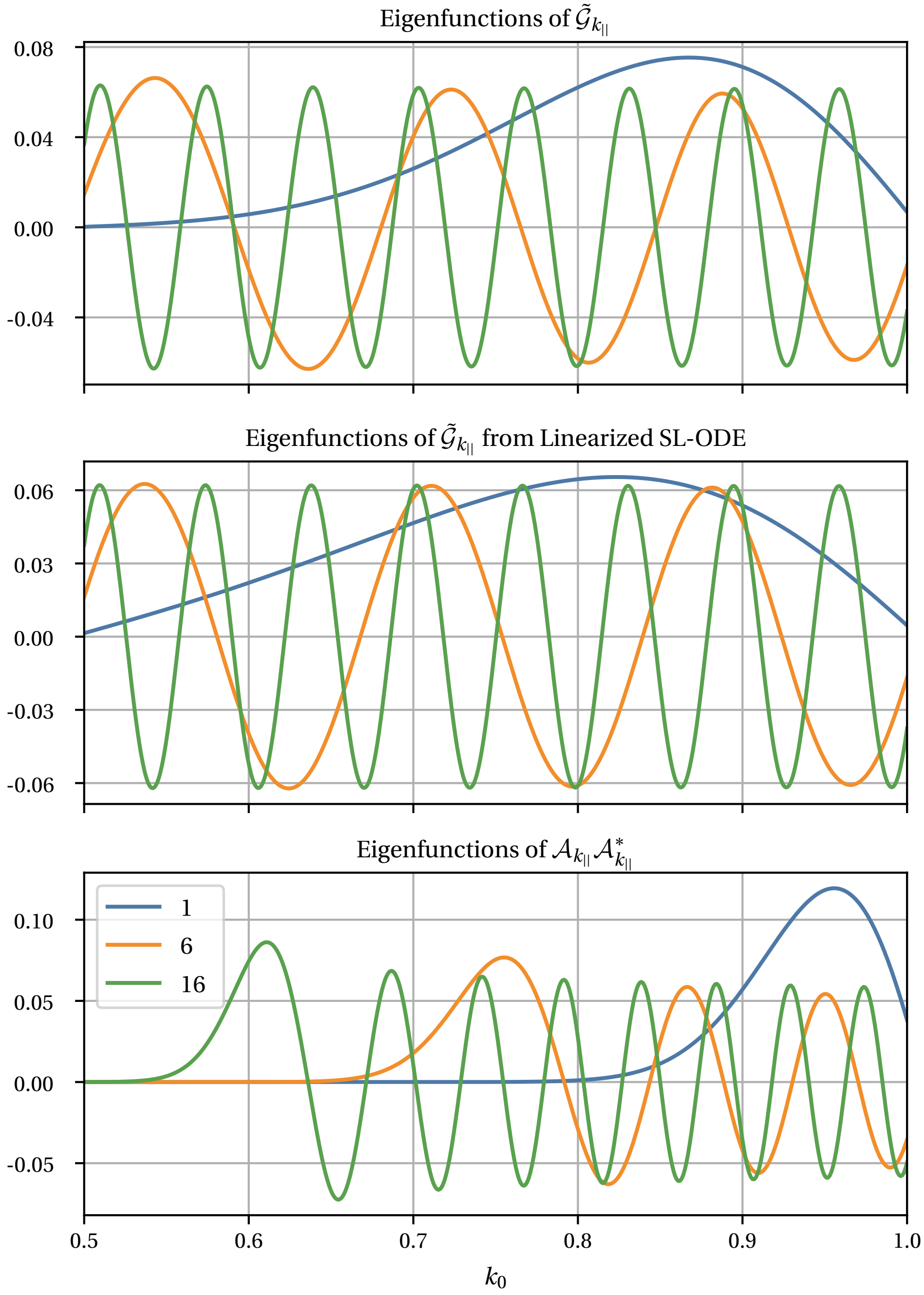


Figure 4.11: Selected eigenfunctions of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ computed according to (4.14), (4.20), and $\mathcal{A}_{\mathbf{k}_\parallel}\mathcal{A}^*_{\mathbf{k}_\parallel}$.

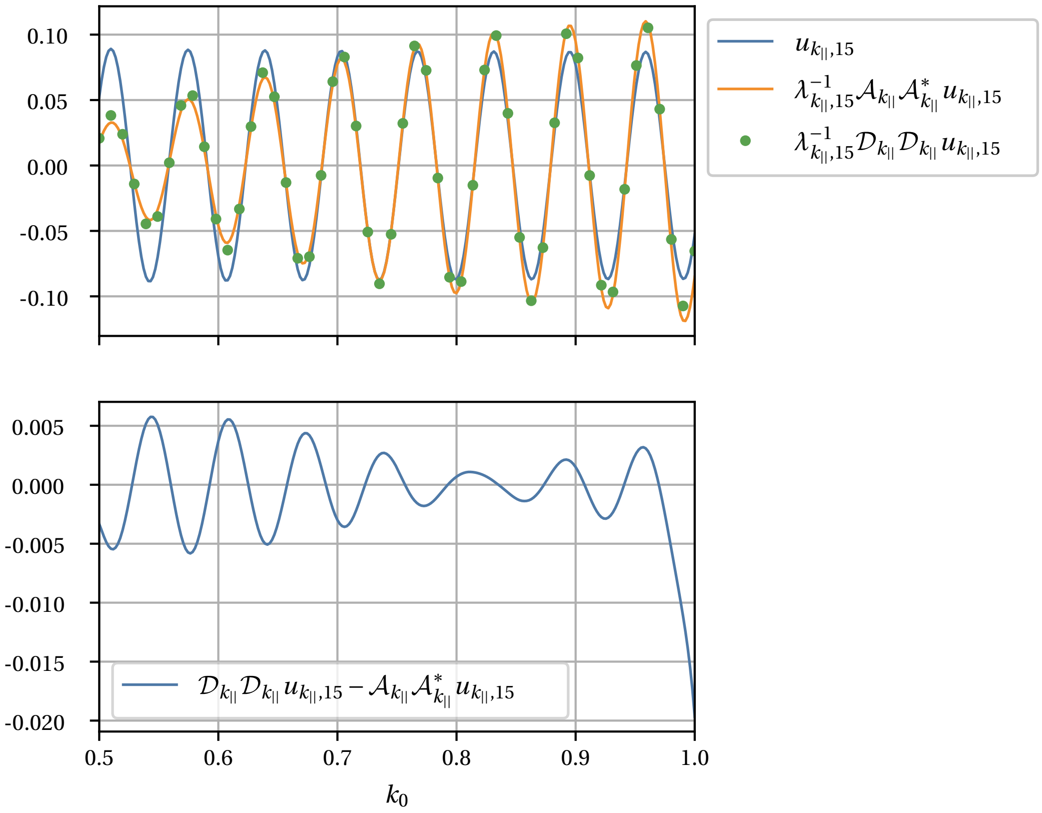


Figure 4.12: (4.12) Demonstrating the generalized eigenfunction relationship (4.12) with $\mathbf{k}_{\|} = (0.15, 0)$.

Algorithm 4 Application of the Forward Operator Using Rank-r Approximation

```
INPUT: $\hat{\eta}(\mathbf{k}_\parallel, z)$
OUTPUT: $\varphi(\mathbf{k}_\parallel, k_0) = (\tilde{\mathcal{A}}\hat{\eta})(\mathbf{k}_\parallel, k_0)$
for $i \leftarrow 1, r$ do
    $\xi_j(\mathbf{k}_\parallel, z) \leftarrow \zeta_j(z)\hat{\eta}(\mathbf{k}_\parallel, z)$    ▷ Depth weighting
    Zero-pad $\xi_j(\mathbf{k}_\parallel, z)$
    $\varphi'_j(\mathbf{k}_\parallel, k_z) \leftarrow \int \xi_j(\mathbf{k}_\parallel, z)e^{-\mathrm{i}k_z z}\mathrm{d}z$    ▷ Apply 1D FFT
    $\varphi_j(\mathbf{k}_\parallel, k_0) \leftarrow w_j(k_0)\varphi'(\mathbf{k}_\parallel, k_z = \phi(\mathbf{k}_\parallel, k_0))$    ▷ Interpolate and Scale
end for
$\varphi(\mathbf{k}_\parallel, k_0) \leftarrow \sum_{j=1}^{r} \varphi_j(\mathbf{k}_\parallel, k_0)$
```

Consider the (unapproximated) ISAM operator, with action $\mathcal{A} : L^2(\mathbb{R}^3) \times L^2(\mathbb{R}^3)$

$$(\mathcal{A}\eta)(\mathbf{k}_\parallel, k_0) = \int_{-\infty}^{\infty} \hat{A}(\mathbf{k}_\parallel, k_0, z)\eta(\mathbf{k}_\parallel, z)\mathrm{d}z. \tag{4.24}$$

We discussed the discretization of (4.24) into a matrix-vector product in Section 3.7.2. Suppose we discretize the spatial domain to N points along each of the x, y and z dimensions, and suppose obtain N wavenumber measurements. Let $\hat{\boldsymbol{\eta}} \in \mathbb{C}^{N^3}$ be the discretization of η; we take $\hat{\boldsymbol{\eta}}$ to be in the transverse Fourier domain. Let $\hat{\mathbf{A}} \in \mathbb{C}^{N^3 \times N^3}$ be the matrix such that $\hat{\mathbf{A}}\hat{\boldsymbol{\eta}}$ is the discretized form of $\mathcal{A}f$. As we discussed, the transverse shift invariance of the ISAM operator leads to a block-diagonal matrix in the transverse (discrete) Fourier domain. Thus $\hat{\mathbf{A}}\hat{\boldsymbol{\eta}}$ can be implemented as N^2 separate dense matrix-vector multiplications; one for each of the N^2 transverse Fourier modes. Each of these matrices is of size $N \times N$. We must store N^4 elements, and applying the forward operator will require $O(N^4)$ FLOPS. If $\hat{\boldsymbol{\eta}}$ is not already in the transverse Fourier domain, the forward operator requires an additional 2D-FFT, but this does not change the storage requirements or the order of computation required.

While the block-diagonal structure of $\hat{\mathbf{A}}$ is helpful (compare to storing N^6 elements and $O(N^6)$ FLOPS), it remains a computationally challenging problem. For a modest problem size of $N = 256$ the direct matrix-vector approach requires ≈ 34 GB of storage (assuming 32-bit complex numbers), although storage needs can be reduced somewhat by exploiting circular symmetry. Moreover, each matrix element requires evaluating a two-dimensional oscillatory integral, and thus cannot be easily computed on-the-fly.

In contrast, our proposed method requires only $O(N)$ storage and applying the forward operator requires $O(N^3 \log N)$ FLOPS.

Algorithm 5 Application of the Adjoint Operator Using Rank-r Approximation

```
INPUT: $\varphi(\mathbf{k}_\parallel, k_0)$
OUTPUT: $f(\mathbf{k}_\parallel, z) = (\tilde{\mathcal{A}}^* \varphi)(\mathbf{k}_\parallel, z)$
for $i \leftarrow 1, r$ do
    $\varrho_j(\mathbf{k}_\parallel, k_0) \leftarrow H(\mathbf{k}_\parallel, k_0)\, w_j(k_0) \varphi(\mathbf{k}_\parallel, k_0)$        ▷ Scale
    $\varrho'_j(\mathbf{k}_\parallel, k_z) \leftarrow \varrho'_j(\mathbf{k}_\parallel, k_z = \phi(\mathbf{k}_\parallel, k_0))$        ▷ Anterpolate
    $f_j(\mathbf{k}_\parallel, z) \leftarrow \zeta_j(z) \int \varrho'_j(\mathbf{k}_\parallel, k_z) e^{\mathrm{i}k_z z} \mathrm{d}z$        ▷ Apply 1D Inverse FFT and Scale
end for
$f(\mathbf{k}_\parallel, z) \leftarrow \sum_{j=1}^{r} f_j(\mathbf{k}_\parallel, z)$
```

4.4.2 Application: Multifocal ISAM

We consider the task of image reconstruction from ISAM data acquired at multiple focal planes. As discussed in Chapter 3, ISAM provides depth-invariant resolution.

By solving the linearized scattering problem, ISAM obtains depth-invariant resolution; however, the signal-to-noise ratio (SNR) can be expected to fall off like $|z - z_F|^{-1}$, ultimately limiting the depth of field. This can be seen as an effect of the function $\zeta_a(z - z_F)$ in the asymptotic approximation (3.11) [152] .

To obtain higher SNR and larger depth of field, Yang *et al.* proposed an extension of ISAM that utilizes data from $N_F \geq 1$ focal planes [153]. Their algorithm initially treats the data from each focal plane as independent and applies the Fourier inversion algorithm. The final reconstruction is the weighted average of the N_F independent reconstructions.

We propose a different method: we pose image reconstruction as the solution of a penalized least squares problem. The resulting optimization problem must be solved using iterative methods, even when Tikhonov regularization is used. We demonstrate that using asymptotic approximate kernel (3.6) in combination with multiple focal planes and an iterative algorithm leads to catastrophic errors.

Let $\{\hat{\mathbf{s}}_{n_F}\}_{n_F=1}^{N_F}$ be the collection of ISAM measurements, where $\hat{\mathbf{s}}_{n_F}$ corresponds to the n_F-th focal plane.

Similarly, $\{\hat{\mathbf{A}}_{n_F}\}_{i=1}^{N_F}$ is the collection of ISAM matrices. At this point, we do not specify how products with $\hat{\mathbf{A}}_{n_F}$ are computed; we will consider both dense matrix multiplication using the unapproximated kernel, and Algorithms 4 and 5. We stack the $\hat{\mathbf{s}}_{n_F}$ and $\hat{\mathbf{A}}_{n_F}$ into $\bar{\mathbf{s}} \in \mathbb{C}^{N_x N_y N_k N_F}$ and $\bar{\mathbf{A}} \in \mathbb{C}^{N_x N_y N_k N_F \times N_x N_y N_z}$.

We want to solve the Tikhonov-regularized least squares problem,

$$\min_{\hat{\boldsymbol{\eta}}} \frac{1}{2} \|\bar{\mathbf{s}} - \bar{\mathbf{A}}\hat{\boldsymbol{\eta}}\|_2^2 + \lambda_r \|\hat{\boldsymbol{\eta}}\|_2^2,$$

Table 4.2: Parameters for image reconstruction examples.

N_x	512	L_x	512.0 µm	Δ_x	1.0 µm
N_z	1536	L_z	1536.0 µm	Δ_z	1.0 µm
N_k	512	k_a	$0.5\ \mathrm{rad}\cdot\mu\mathrm{m}^{-1}$	k_b	$1.0\ \mathrm{rad}\cdot\mu\mathrm{m}^{-1}$
NA	0.2	$\lambda_{\min}$	$6.3\ \mu\mathrm{m}^{-1}$	$\lambda_{\max}$	$12.6\ \mu\mathrm{m}^{-1}$
N_F	2	z_F	$[-225, 225]$ µm		

where $\|\cdot\|_2$ is the ℓ_2 norm and $\lambda_r > 0$ is the scalar regularization parameter.

The normal equations are

$$\left(\bar{\mathbf{A}}^{\mathsf{H}}\bar{\mathbf{A}} + \lambda_r \mathbf{I}\right)\hat{\boldsymbol{\eta}} = \bar{\mathbf{A}}^{\mathsf{H}}\bar{\mathbf{s}},$$

or in expanded form,

$$\left(\sum_{n_F=1}^{N_F} \hat{\mathbf{A}}_{n_F}^{\mathsf{H}}\hat{\mathbf{A}}_{n_F} + \lambda_r \mathbf{I}\right)\hat{\boldsymbol{\eta}} = \sum_{n_F=1}^{N_F} \hat{\mathbf{A}}_{n_F}^{\mathsf{H}}\hat{\mathbf{s}}_{n_F}. \tag{4.25}$$

We solve the linear system (4.25) using the Conjugate Gradient (CG) algorithm. CG requires only products with $\bar{\mathbf{A}}$ and $\bar{\mathbf{A}}^{\mathsf{H}}$.

We consider three methods to apply the ISAM forward operator. First, we use dense matrix multiplication, where $\hat{\mathbf{A}}_{n_F}$ is the discretized version of the exact ISAM kernel as described in Section 3.7.2. We refer to this as the "exact approach". Second, we consider the use of the asymptotic kernel (3.6). To reconcile the coupling between k_0 and z, as discussed in Section 3.6.1, we use the hybrid approach (3.11). In this case, the forward and adjoint operators are applied using Algorithms 4 and 5. We call this the "asymptotic approach". Finally, we use the proposed low-rank factorization of $W(k_0, z)$ given by (4.7) in combination with Algorithms 4 and 5. We call this the "rank-r" approach.

We implement our algorithms on a NVidia 1080 Ti GPU using a combination of Python and CUDA [87,88].

4.4.3 Experiments

Our simulation is restricted to two spatial dimensions, (x, z). Our phantom consists of 16 point scatterers lying on the $x = 0$ spaced 50 µm apart, covering a total of 750 µm. We use two focal planes, located at ±225 µm. The system remaining system parameters are listed in Table 4.2.

We generated data according to the exact forward model (3.9). We model the point scatterers as a sum of Dirac delta functions,

$$\begin{aligned}\eta(\mathbf{r}_\parallel, z) &= \sum_j \delta^{(3)}\left(\mathbf{r}_\parallel - \mathbf{r}_{||,j}, z - z_j\right)\\ \hat{\eta}(\mathbf{k}_\parallel, z) &= \sum_j e^{\mathrm{i}\mathbf{r}_{||,j}\cdot\mathbf{k}_\parallel}\delta(z - z_j).\end{aligned}$$

Inserting this form into (3.9) and applying the sifting property of the delta function yields our measurements,

$$\begin{aligned}\hat{S}(\mathbf{k}_\parallel, z_F, k_0) &= \int \hat{A}(\mathbf{k}_\parallel, z - z_F, k_0) \sum_j e^{\mathrm{i}\mathbf{r}_{||,j}\cdot\mathbf{k}_\parallel}\delta(z - z_j)\mathrm{d}z\\ &= \sum_j \hat{A}(\mathbf{k}_\parallel, z_j - z_F, k_0)e^{\mathrm{i}\mathbf{r}_{||,j}\cdot\mathbf{k}_\parallel}.\end{aligned}$$

We obtain $\hat{A}(\mathbf{k}_\parallel, z_j - z_F, k_0)$ at the necessary locations by evaluating (3.4) using numerical quadrature.

We applied 1000 CG iterations with $\lambda_r = 300$. The magnitude of the reconstructed images are shown in Fig. 4.13. Figure 4.13(a) shows the result using the "exact" approach. This serves as our baseline for further comparisons. The point targets are correctly localized and they appear to have uniform magnitude. This is validated in Fig. 4.14, where we plot the profile of the reconstruction along $x = 0$. For clarity, we show only $z > 0$. The "exact" reconstruction shows some amplitude nonuniformity near the focal plane ($z = 225$). Figure 4.13(b) shows the result using the "asymptotic" approach. This reconstruction is clearly incorrect. The bright artifacts occur near $z = \pm 170$; this is where we switch from the near-focus to far-from-focus regimes. We do not include these results in Fig. 4.14.

Figure 4.13(c) shows the results using the rank-r approach with $r = 1$; this is our proposed modification of the narrowband and paraxial approximation from the original ISAM paper [99]. The results are much better than the asymptotic case, but upon examining the 1D profiles in Fig. 4.14, we see the amplitude fluctuates up to 20%. Note that in this experiment, $k_a = 0.5$ and $k_b = 1$. This is not a good fit for the narrowband approximation, which holds for $(k_b - k_a)/k_a \approx 1$.

Finally, Fig. 4.13(d) shows the result using the rank-r approach with $r = 2$. The results are in close agreement with the "exact" case. This demonstrates that a higher-order approximation to W is useful in extending the approximation beyond the narrowband regime.

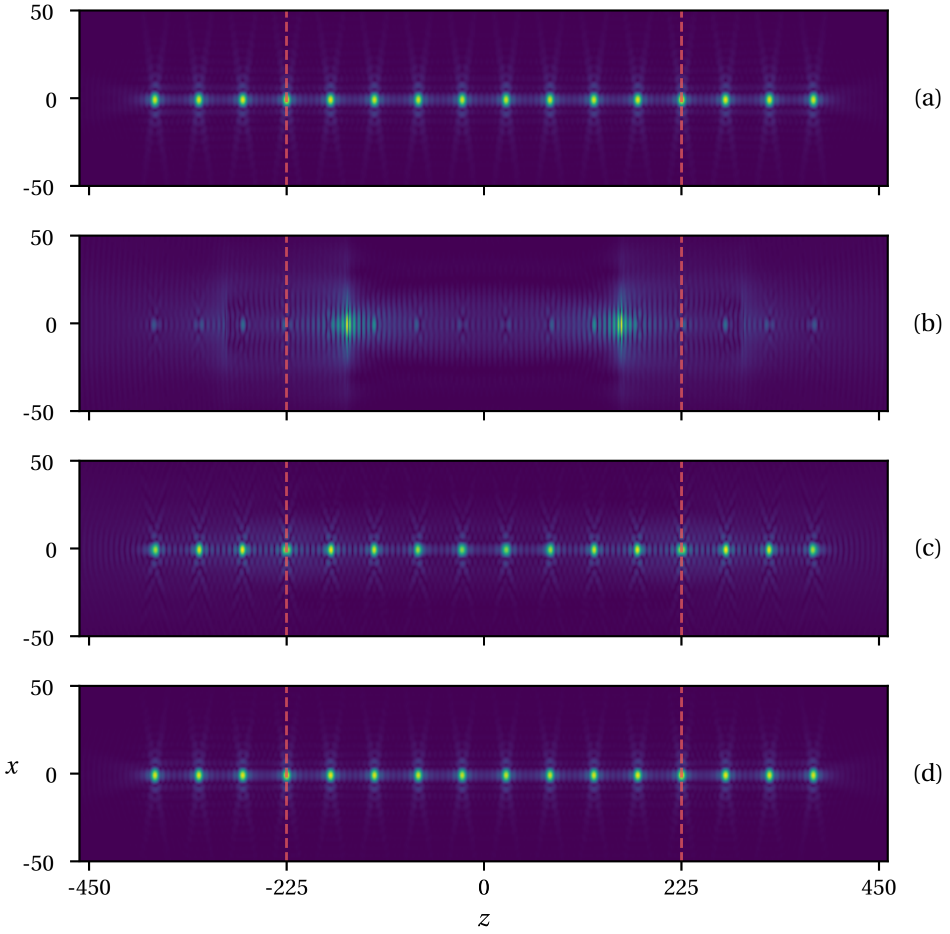


Figure 4.13: Modulus of reconstructed point targets from data acquired at two focal planes. System parameters listed Table 4.2. The red dashed lines denote the location of the two focal planes. All reconstructions used Tikhonov regularization with $\lambda = 10^2$. (a) Reconstruction using the exact ISAM kernel. (b) Reconstruction using the asymptotic approximate ISAM kernel described in Section 3.4.2. (c) Reconstruction using the rank-one approximation (4.4) with one term ($r = 1$). (d) Reconstruction using the rank-r approximation (4.7) with two terms ($r = 2$).

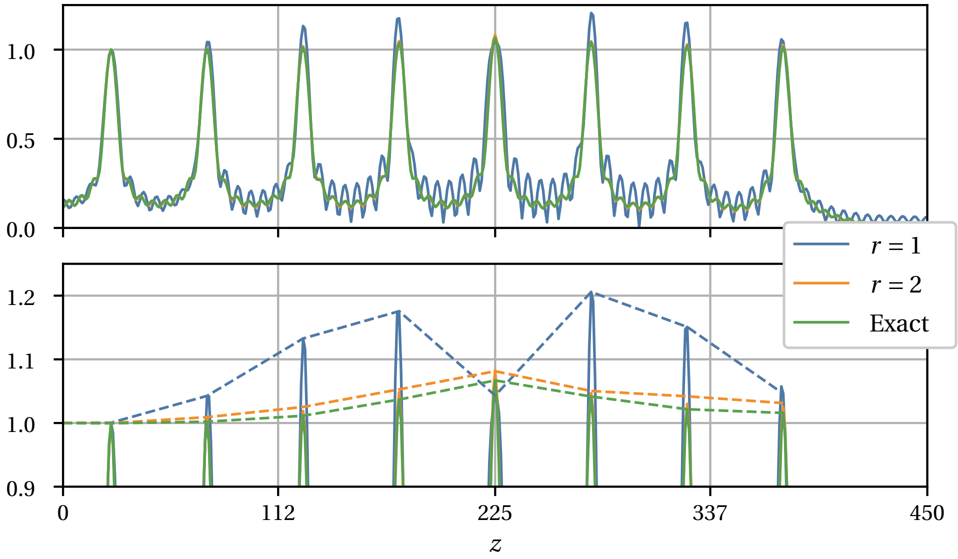


Figure 4.14: Top : Horizontal profiles of the modulus of reconstructed point targets shown in Fig. 4.13. Profiles taken at $x = 0$. For clarity, we show only the halfspace $z > 0$. Bottom: Zooming in to the behavior of the reconstructed peaks. The dashed lines follow connect the reconstructed peaks to highlight the difference between the reconstructions with $r = 2$ and the exact kernel.

Chapter 5

Composition-Aware Spectroscopic Tomography

5.1 Introduction

Chemically specific imaging provides quantitative information about the distribution of chemicals within a target. This may be accomplished through the use of exogenous chemicals or molecular staining to improve contrast when the target is imaged with visible light. For many applications, these application of these dyes cannot be introduced *in situ*, and the agents are often damaging to the target.

Vibrational spectroscopy with mid-infrared light presents a solution [154]. Absorption of mid-infrared light depends on chemical composition. The underlying chemistry of a target can be determined, non-invasively, by illuminating the object with mid-infrared light and recording an absorption spectrum.

In principle, mid-infrared spectroscopy can provide chemically specific, spatially resolved imaging in three spatial dimensions using a confocal scanning strategy: the target would be scanned point-by-point in three spatial dimensions, and an absorption spectrum would be measured at each point [155, 156]. For a target with two spatial dimensions, this is feasible- a typical data set of 1024 spectral samples over a 1024 × 1024 pixel grid requires on the order an of hour of acquisition time and generates roughly 25 GB of data. Scanning along a third spatial dimension (depth) makes imaging even a single target impractical: the resulting dataset would require over 25 terabytes of storage and roughly a month of acquisition time.

The key challenge in jointly measuring structural and chemical information is dimensionality: with no constraints, the target can vary in three spatial and one spectral dimension. Existing imaging modalities explicitly or implicitly rely on simple signal models to reduce the dimensionality of the target and allow for practical imaging.

Optical Coherence Tomography (OCT) and Interferometric Synthetic Aperture Microscopy (ISAM) are scattering-based imaging modalities that reconstruct the 3D spatial distribution of a target by ignoring spectral variation, although limited spectral information can be recovered at the expense of spatial resolution by way of time-frequency analysis [157–159].

Fourier Transform Infrared (FTIR) spectroscopy, a workhorse of academic and industrial labs worldwide, neglects all spatial variation within the target—thus reducing the target to a single dimension. An extension, FTIR microspectroscopy, provides spatially and spectrally resolved measurements but requires the target to be very thin with only transverse heterogeneities. Unmodeled spatial variations in the target cause scattering and diffraction, ultimately distorting the measured spectra [155, 156].

We propose an approach that bridges these two extremes and allows for practical, chemically specific imaging. We call this *spectroscopic tomography*. Rather than finely scanning the focus through the axial dimension of the target, we acquire data at a small number of *en-face* focal planes. The target is recovered by solving the linearized scattering problem. A low-dimensional model is used to regularize the inverse problem: we model the target as the linear combination of a finite number of distinct chemical species. This is called the *N-species* approximation. We develop a set of algebraic conditions for unique recovery and examine the conditioning of the inverse problem. Reconstructions from synthetic phantom data illustrate the promise of the model.

Preliminary research in this direction considered this problem, and the N-species model, with sample variation in one spatial dimension [160]. Their simulated results involve several unrealistic assumptions, leading to results of unrealistically high quality. We extend this work in several directions: we (i) use a non-asymptotic forward model; (ii) demonstrate material-resolved reconstruction of samples with two spatial dimensions (one transverse and depth, easily extended to three spatial dimensions) from data that is not generated according to the first Born approximation; and (iii) refine the conditions for recovery of a sample consisting of N-species from interferometric scattering experiments.

In Section 5.2 we describe our measurement model. Section 5.3 describes the N-species model in greater detail. We discuss the sampling and discretization procedure in Section 5.4. We investigate the inverse problem in Section 5.5, and demonstrate the method by performing numerical reconstructions from simulated measurements in Section 5.6.

5.1.1 Notation

We write the set of integers $\{1, 2, \ldots, N\}$ as $[N]$ and the imaginary unit as i. Finite-dimensional vectors are denoted by lower-case bold letters, *e.g.* $\mathbf{x} \in \mathbb{C}^N$. Finite-dimensional matrices and tensors are written using upper-case bold letters. We adopt Matlab-style indexing notation: given a matrix $\mathbf{A} \in \mathbb{C}^{N \times M}$, its i-th row is $\mathbf{A}[i, :\,]$, the j-th column is $\mathbf{A}[:\,, j]$, and i, j-th element is $\mathbf{A}[i, j]$. We denote the vector $\operatorname{vec}(\mathbf{A}) \in \mathbb{C}^{NM}$ is formed by stacking the columns of $\mathbf{A}$ into a single vector (*i.e.*, row-major ordering). The range, null space, and rank of a matrix $\mathbf{A}$ are written range $\{\mathbf{A}\}$, null $\{\mathbf{A}\}$, and rank $\{\mathbf{A}\}$. Given $\mathbf{x} \in \mathbb{C}^N$, the diagonal matrix diag $\{\mathbf{x}\} \in \mathbb{C}^{N\times N}$ has the entries of $\mathbf{x}$ along its main diagonal. Similarly, given a set of $N \times M$ matrices $\mathbf{A}_1, \ldots, \mathbf{A}_L$, the matrix blkdiag $(\mathbf{A}_1, \ldots, \mathbf{A}_L) \in \mathbb{C}^{LN\times LM}$ is block-diagonal with the collection of $\mathbf{A}_i$ along its block diagonal.

The transpose (*resp.* Hermitian transpose) of a matrix is written $\mathbf{A}^\mathsf{T}$ (*resp.* $\mathbf{A}^\mathsf{H}$). The ℓ_p norm of $\mathbf{x} \in \mathbb{C}^N$ is $\|\mathbf{x}\|_p = \left(\sum_{j=1}^N \left|\mathbf{x}[j]\right|^p\right)^{1/p}$. For vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$ we use the shorthand $|r| = \|\mathbf{r}\|_2$. The $N \times N$ identity matrix is $\mathbf{I}_N$, and the vector $[1, 1, \ldots 1]^\mathsf{T} \in \mathbb{R}^N$ is written $\mathbb{1}_N$. The tensor (or Kronecker) product between matrices $\mathbf{A}$ and $\mathbf{B}$ is $\mathbf{A} \otimes \mathbf{B}$.

5.2 Forward Model: Interferometric Synthetic Aperture Microscopy

Our forward model is described in detail in Chapter 3. To review, we model our sample through its complex susceptibility $\eta(\mathbf{r}_\parallel, k_0)$. Here, $\mathbf{r} = (x, y, z) = (\mathbf{r}_\parallel, z)$, where $\mathbf{r}_\parallel$ are the transverse dimensions and z indicates the axial dimension. We assume that η is (spatially) supported in the bounded region $\Gamma \subset \mathbb{R}^3$. The free-space wavenumber is k_0. Importantly, we do not assume the object is non-dispersive.

In the transverse Fourier domain, our measurements are of the form

$$\hat{S}(\mathbf{k}_\parallel, z_F, k_0) = \int \hat{A}(\mathbf{k}_\parallel, z - z_F, k_0)\hat{\eta}(\mathbf{k}_\parallel, z, k_0)\, dz, \tag{5.1}$$

where $\hat{A}$ is the *unapproximated* ISAM kernel given by (3.4). We do not use the approximate formulations described in Section 3.4 or Chapter 4. However, we use the insight provided by these approximations as a guide; in particular, the projection-slice interpretation described in Section 3.5 informs our sampling procedure and helps to establish fundamental limits of the imaging system.

5.3 The N-Species Model

5.3.1 The Model

The fundamental problem of spectroscopic tomography is the dimensionality of the sample: an arbitrary sample can vary in four dimensions (three spatial and one spectral). As discussed in Section 3.5, measurements of the form (5.1) acquired at a single focal plane can be related to the Fourier transform of the sample along a three-dimensional surface. Acquiring a fourth dimension of data—in our case, by scanning in three spatial and one spectral dimension—is prohibitively expensive.

Existing imaging modalities use simplified signal models to reduce the dimensionality of the sample and allow for practical imaging. We have seen that ISAM assumes is either non-dispersive or has (known) spatially invariant dispersion characteristics. In this case, the susceptibility is of the form $\eta(\mathbf{r}, k_0) = p(\mathbf{r})h(k_0)$, where $p(\mathbf{r})$ captures the spatial density of the target and $h(k_0)$ characterizes the wavelength-dependent dispersion characteristics. If $h(k_0)$ is known, only $p(\mathbf{r})$ must be determined—thus reducing the problem to recovery of a three-dimensional object. Diffraction tomography, reflection tomography, and optical coherence tomography also assume non-dispersive targets. Conversely, Fourier Transform InfraRed (FTIR) spectroscopy of a bulk medium assumes that the sample is spatially homogeneous, so that $\eta(\mathbf{r}, k_0) = h(k_0)$. An extension, FTIR microscopy, models the sample as a thin absorbing screen; thus $\eta(\mathbf{r}, k_0) = \eta(\mathbf{r}_\parallel, k_0)$, a three-dimensional object.

These examples severely restrict the class of samples that can be imaged. We propose a model that is more expressive than these examples while still allowing practical imaging.

Definition 5.1 (The N-species model [160])**.** *An object, described by a susceptibility $\eta(\mathbf{r}, k_0)$, is said to satisfy the N-species model if*

$$\eta(\mathbf{r}, k_0) = \sum_{n_s=1}^{N_s} p_{n_s}(\mathbf{r}) h_{n_s}(k_0). \tag{5.2}$$

The function $p_{n_s}(\mathbf{r})$ captures the spatial variation of the n_s-th species and is called the spatial density. *If species n_s is not present at location $\mathbf{r}$, then $p_{n_s}(\mathbf{r}) = 0$. The complex function h_{n_s} models the wavelength-dependent properties of the n_s-th species and is called the* spectral profile.

The N-species model, introduced in [160], is a rank N_s approximation to a general susceptibility. A similar decomposition has been applied to magnetic resonance spectroscopic imaging; in this context, it is called the Partially Separable (PS) function model [161–164]. A similar model is used for material decomposition in X-ray tomography [165, 166].

The N-species model (5.2) is physically justifiable. The susceptibility of a linear, isotropic, dielectric medium is a well-modeled by a sum of Lorentzian functions [167]; that is,

$$\eta(\mathbf{r}, k_0) = p(\mathbf{r}) \left(\sigma_0(\mathbf{r}) + \sum_{n=1}^{N_l - 1} \frac{\sigma_n(\mathbf{r})}{\nu_n^2(\mathbf{r}) - k_0^2 - \mathrm{i}\gamma_n(\mathbf{r}) k_0} \right), \tag{5.3}$$

where N_l is the number of Lorentzian functions, ν_n is called the resonance frequency, σ_n is the oscillator strength and γ_n is the damping constant. These quantities depend on the electron binding characteristics of the medium. The spatial density is defined as $p(\mathbf{r}) \triangleq \varrho(\mathbf{r}) e^2 (m\epsilon_0)^{-1}$, where $\varrho(\mathbf{r})$ is the electron density of the medium, ϵ_0 is the permittivity of free space, and e, m are the electron charge and mass, respectively.

A single-species model, based on (5.3), has been used to compensate for dispersion effects in multispectral intensity-only diffraction tomography [168]. It is assumed that the electron binding characteristics of the medium are uniform within the medium; only the electron density ϱ is position dependent. We extend this to the multiple-species regime. Suppose the object η comprises N_s distinct materials. For $n_s \in [N_s]$, let ϱ_{n_s} be the electron density of the n_s-th chemical, and define

$$p_{n_s}(\mathbf{r}) \triangleq \begin{cases} \frac{\varrho_{n_s}(\mathbf{r}) e^2}{m\epsilon_0}, & \text{if species } n_s \text{ present at location } \mathbf{r}, \\ 0, & \text{otherwise.} \end{cases}$$

As the electron binding characteristics are spatially invariant by assumption, we omit the spatial dependence of the parameters $\sigma_{n_s,n}, \gamma_{n_s,n}$, and $\nu_{n_s,n}$; here, the subscript n_s, n associates the parameter with n-th Lorentzian function for the n_s-th material. By defining

$$h_{n_s}(k_0) \triangleq \sigma_{n_s,0} + \sum_{n=1}^{N_{l,n_s}-1} \frac{\sigma_{n_s,n}}{\nu_{n_s,n}^2 - k_0^2 - \mathrm{i}\gamma_{n_s,n} k_0}, \tag{5.4}$$

we can write the complete susceptibility in the desired form (5.2).

5.3.2 Spectroscopic Tomography with the N-Species Model

Inserting the N-species model (5.2) into the linearized forward model (3.3), we have

$$\hat{S}(\mathbf{k}_\parallel, k_0, z_F) = \sum_{n_s=1}^{N_s} h_{n_s}(k_0) \int_{-\infty}^{\infty} \hat{A}(\mathbf{k}_\parallel, z - z_F, k_0) \hat{p}_{n_s}(\mathbf{k}_\parallel, z) \mathrm{d}z. \tag{5.5}$$

At a given focal plane, the measurements are the sum of N_s independent ISAM experiments, each on a non-dispersive object $\hat{p}_{n_s}(\mathbf{k}_\parallel, z)$ and each weighted by the spectral profile $h_{n_s}(k_0)$. In what follows, we study inverse problem associated with spectroscopic optical tomography: we wish to recover an object that satisfies the N-species model from measurements of the form (5.5).

We know that in the single species case, the inverse problem can be solved from data acquired at single focal plane—this is the usual ISAM problem. On the other hand, an arbitrary sample can be recovered by finely scanning in all three spatial dimensions (*i.e.*, along $\mathbf{r}_\parallel^{(o)}$ and z_F) and acquiring a spectrum at each point, but this is infeasible as described in Section 5.1.

The N-species model is a middle ground between a single species object and an arbitrary one. Our goal is to show that the number of measurements required to solve the inverse problem also lies in a middle ground between these two extremes: in particular, we hope that an object satisfying the N-species model can be recovered using $N_F \approx N_s$ focal planes.

We divide the inverse problem into three distinct cases.

(P1) *Known Spectra.* Assume the spectral profiles $\{h_{n_s}\}_{n_s=1}^{N_s}$ are fixed and known. Our task reduces to a linear inverse problem—recovery of the $\{\hat{p}_{n_s}\}_{n_s=1}^{N_s}$ from measurements of the form (5.5).

(P2) *Spectra from a Dictionary.* Assume the target comprises at most N_s chemical species, but the spectral profiles are drawn from a (known) dictionary of some $M_s > N_s$ possible spectra. The inverse problem can be phrased as either a linear inverse problem over the entire dictionary, or as a nonlinear problem where the solution is constrained to lie in a union of subspaces.

(P3) *Fully Blind.* Both the $\{h_{n_s}\}_{n_s=1}^{N_s}$ and $\{\hat{p}_{n_s}\}_{n_s=1}^{N_s}$ are unknown and must be recovered from measurements of the form (5.5). This is a *bilinear* inverse problem in h_{n_s} and p_{n_s}.

In this chapter, we limit our attention to cases (P1) and (P2). Our analysis is based on a discretized form of (5.5) wherein all quantities are replaced by finite-dimensional versions, resulting in a so-called "discrete-to-discrete" inverse problem [107, 117]. Next, we describe our sampling and discretization procedure.

5.4 Sampling and Discretization of the Forward Model

5.4.1 Sampling

The instrument acquires samples of the spatial-domain measurement equation (3.2). We assume the object is (spatially) supported in a region $\Gamma \subset \mathbb{R}^3$; here, we take $\Gamma = [0, L_x] \times [0, L_y] \times [0, L_z]$. We write the number of samples as N_i and the discretization or sampling interval as Δ_i for $i = x, y, z, k$. We obtain measurements at the transverse aperture locations $\mathbf{r}_{\parallel}^{(o)} = (n_x\Delta_x, n_y\Delta_y)$ for integers n_x, n_y. The parameters are chosen to cover Γ, *i.e.* $N_i\Delta_i = L_i$ holds for $i = x, y, z$. For simplicity, we assume the sampling parameters are the same along the x and y directions: $N_x = N_y$, $\Delta_x = \Delta_y$, and $L_x = L_y = N_x\Delta_x$. The wavenumber is sampled uniformly over the interval $[k_a, k_b]$ with sampling interval Δ_k; the n_k-th measurement wavenumber is $k_{0,i} \triangleq k_a + n_k\Delta_k$. We acquire data at N_F focal planes, written $\{z_{F,i} : i = 1,2,\dots N_F\}$. The same sampling parameters are used at each focal plane; in particular, the set of sampled wavenumbers does not change.

We choose the sampling parameters as we would for a standard, single-species ISAM problem. The necessary sampling intervals can be motivated using the approximate forward model studied in Chapter 4. Under this model, it can be shown that "point spread function" $\left|A(\mathbf{r}_\parallel, k_0, z)\right|$ (approximately) decays like a Gaussian in $\left|\mathbf{r}_\parallel\right|$. We take L_x and L_y large enough to safely neglect the unmeasured data. Moreover, for fixed z_F the measurements $\hat{S}(\mathbf{k}_\parallel, k_0, z_F)$ are bandlimited to $[-k_b \sin \mathrm{NA}, k_b \sin \mathrm{NA}]$; we sample along the transverse dimension at intervals $\Delta_x, \Delta_y < \pi/(k_b \sin \mathrm{NA})$. Finally, the combination of uniform sampling in $\mathbf{r}_\parallel^{(o)}$ and k_0 leads to a non-uniform sampling of the Fourier transform of the object: samples are obtained at uniform locations along the $\mathbf{k}_\parallel$ axis but at nonuniform locations along the k_z axis. To avoid aliasing, we require that the maximum distance between samples on the k_z axis is less than π/L_z [121, 122].

5.4.2 Discretization

Given samples of (3.2), we take the 2D Discrete Fourier Transform (DFT) with respect to the transverse coordinates and write the result as the tensor $\hat{\mathbf{S}} \in \mathbb{C}^{N_x \times N_y \times N_k \times N_F}$. We continue to assume $N_x = N_y$ with N_x an even integer. The 2D-DFT coordinate $\mathbf{q}_\parallel = (q_x, q_y)$ is an integer vector with $0 \le q_x, q_y \le N_x - 1$. We obtain the continuous Fourier coordinate k_x from the DFT coordinate k_x as

$$k_x(q_x) = \begin{cases} 2\pi q_x / L_x & q_x < N_x/2 \\ 2\pi(q_x - N_x)/L_x & \text{otherwise,} \end{cases} \tag{5.6}$$

and the same holds for q_y and k_y. We define $\mathbf{k}_\parallel(\mathbf{q}_\parallel) = \left(k_x(q_x), k_y(q_y)\right)$.

The discretized N-species measurement model is

$$\hat{\mathbf{S}}[\mathbf{q}_\parallel, n_k, n_F] = \sum_{n_s=1}^{N_s} \mathbf{h}_{n_s}[n_k] \sum_{n_z=0}^{N_z-1} \hat{\mathbf{A}}_{n_F}[\mathbf{q}_\parallel, n_k, n_z] \hat{\mathbf{P}}_{n_s}[\mathbf{q}_\parallel, n_z], \tag{5.7}$$

where $\mathbf{h}_{n_s} \in \mathbb{C}^{N_k}$ and $\hat{\mathbf{P}}_{n_s} \in \mathbb{C}^{N_x \times N_y \times N_z}$ are the discretized spectral profile and spatial density corresponding to the n_s-th species, respectively, and

$$\hat{\mathbf{A}}_{n_F}[\mathbf{q}_\parallel, n_k, n_z] \triangleq \hat{A}\left(\mathbf{k}_\parallel(\mathbf{q}_\parallel), k_a + n_k \Delta_k, N_z \Delta_z - z_{F,n_F}\right).$$

Additionally, we gather the Fourier transforms of the discrete spatial densities into $\hat{\mathbf{P}} \in \mathbb{C}^{N_x \times N_y \times N_z \times N_s}$

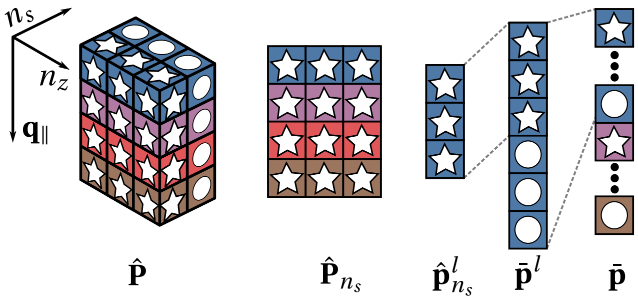


Figure 5.1: The various unfoldings of the discretized spatial densities with $N_s = 2$. Here, block color indicates the value of $\mathbf{q}_\parallel$. Species 1 is marked with a star, while species 2 is indicated with a circle.

and spectral profiles into $\mathbf{H} \in \mathbb{C}^{N_k \times N_s}$, with

$$\hat{\mathbf{P}}[\mathbf{q}_\parallel, n_z, n_s] = \hat{\mathbf{P}}_{n_s}[\mathbf{q}_\parallel, n_z]$$
$$\mathbf{H}[n_k, n_s] = \mathbf{h}_{n_s}[n_k].$$

5.4.3 Block-Matrix Form of N-Species Forward Model

With the spectral profiles fixed, the measurements $\hat{\mathbf{S}}$ are a linear function of the spatial densities. Thus we can write (5.7) as a matrix-vector product, where the vector depends only on the spatial densities. The resulting matrix has a block-diagonal structure which is key to our analysis of the discretized inverse problem.

Exploring this structure requires slicing and reshaping the tensors $\hat{\mathbf{S}}, \hat{\mathbf{A}}_{n_F}$, and $\hat{\mathbf{P}}$ into a variety of forms. We introduce additional notation to represent these derived quantities; the various forms of $\hat{\mathbf{P}}$ are illustrated in Fig. 5.1. Recall upper-case bold letters refer to matrices or tensors and lower-case bold letters refer to vectors. We use a bar to denote objects that have been "stacked" or vectorized. Subscripts are used to slice a tensor with respect to the last index: *e.g.* $\hat{\mathbf{s}}_{n_F}$ represents all measurements from the n_F-th focal plane, while $\mathbf{h}_{n_s}$ and $\hat{\mathbf{P}}_{n_s}$ are the spectral profile and spatial density for the n_s-th species. A superscript indicates a submatrix or vector formed for particular value of $\mathbf{q}_\parallel$. We use the reindexing function

$$\gamma : \mathbb{Z}^2 \to \mathbb{Z} \quad \gamma(\mathbf{q}_\parallel) = q_x + N_x q_y,$$

to identify the 2D-DFT index $\mathbf{q}_\parallel$ with the integer $\gamma(\mathbf{q}_\parallel)$. Let $l = \gamma(\mathbf{q}_\parallel)$ and define

$$\begin{aligned}
\hat{\mathbf{s}}_{n_F}^l &\triangleq \hat{\mathbf{S}}[\gamma^{-1}(l), :, n_F] \in \mathbb{C}^{N_k} \\
\hat{\mathbf{A}}_{n_F}^l &\triangleq \hat{\mathbf{A}}_{n_F}[\gamma^{-1}(l), :, :] \in \mathbb{C}^{N_k \times N_z} \\
\hat{\mathbf{p}}_{n_s}^l &\triangleq \hat{\mathbf{P}}[\gamma^{-1}(l), :, n_s] \in \mathbb{C}^{N_z}.
\end{aligned}$$

Further, define the diagonal matrix $\mathbf{D}_{n_s} \triangleq \operatorname{diag}\left(\mathbf{h}_{n_s}\right) \in \mathbb{C}^{N_k \times N_k}$. Now, for fixed $l = \gamma(\mathbf{q}_\parallel)$ and n_F, (5.7) is equivalent to

$$\hat{\mathbf{s}}_{n_F}^l = \sum_{n_s=1}^{N_s} \mathbf{D}_{n_s} \hat{\mathbf{A}}_{n_F}^l \hat{\mathbf{p}}_{n_s}^l. \tag{5.8}$$

The collection of (5.8) for $n_F \in [N_F]$ can be written as a single linear system. Define the vectors

$$\begin{aligned}
\bar{\mathbf{p}}^l &\triangleq \operatorname{vec}\left(\hat{\mathbf{P}}[\gamma^{-1}(l), :, :]\right) = [(\hat{\mathbf{p}}_1^l)^\top, \dots, (\hat{\mathbf{p}}_{N_s}^l)^\top]^\top \in \mathbb{C}^{N_s N_z} \\
\bar{\mathbf{s}}^l &\triangleq \operatorname{vec}\left(\hat{\mathbf{S}}[\gamma^{-1}(l), :, :]\right) = [(\hat{\mathbf{s}}_1^l)^\top, \dots, (\hat{\mathbf{s}}_{N_F}^l)^\top]^\top \in \mathbb{C}^{N_F N_k},
\end{aligned}$$

which contain the spatial densities for each species and measurements for all focal planes, respectively, and the block matrix $\Phi^l \in \mathbb{C}^{N_F N_k \times N_s N_z}$ by

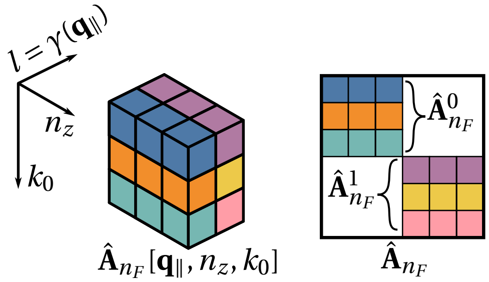
$$\Phi^l \triangleq \begin{bmatrix} \mathbf{D}_1 \hat{\mathbf{A}}_1^l & \dots & \mathbf{D}_{N_s} \hat{\mathbf{A}}_1^l \\ \vdots & \ddots & \vdots \\ \mathbf{D}_1 \hat{\mathbf{A}}_{N_F}^l & \dots & \mathbf{D}_{N_s} \hat{\mathbf{A}}_{N_F}^l \end{bmatrix}. \tag{5.9}$$

Each block-row of Φ^l corresponds to the $l = \gamma(\mathbf{q}_\parallel)$ transverse Fourier component of measurements taken at a single focal plane, and the n_s-th block-column corresponds to the n_s-th species. With these definitions in place, we have

$$\bar{\mathbf{s}}^l = \Phi^l \bar{\mathbf{p}}^l. \tag{5.10}$$

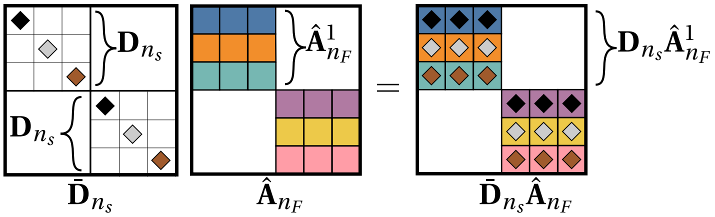
Equation (5.10) is the discretized N-species forward model at a single transverse Fourier frequency $\mathbf{q}_\parallel = \gamma^{-1}(l)$.

We can form an analogous linear system that describes the forward model for all $\mathbf{q}_\parallel$. We stack



$l = \gamma(\mathbf{q}_\parallel)$
n_z
k_0
$\hat{\mathbf{A}}_{n_F}[\mathbf{q}_\parallel, n_z, k_0]$
$\hat{\mathbf{A}}^0_{n_F}$
$\hat{\mathbf{A}}^1_{n_F}$
$\hat{\mathbf{A}}_{n_F}$

Figure 5.2: Unfolding the tensor $\hat{\mathbf{A}}_{n_F}[\mathbf{q}_\parallel, n_z, n_k]$ into a block-diagonal matrix.



$\mathbf{D}_{n_s}$
$\mathbf{D}_{n_s}$
$\bar{\mathbf{D}}_{n_s}$
$\hat{\mathbf{A}}^1_{n_F}$
$\hat{\mathbf{A}}_{n_F}$
$=$
$\mathbf{D}_{n_s}\hat{\mathbf{A}}^1_{n_F}$
$\bar{\mathbf{D}}_{n_s}\hat{\mathbf{A}}_{n_F}$

Figure 5.3: Constructing the scaled ISAM matrices at a single focal plane.

the $\{\bar{\mathbf{p}}^l\}_{l=0}^{N_xN_y-1}$ and $\{\bar{\mathbf{s}}^l\}_{l=0}^{N_xN_y-1}$ into vectors $\bar{\mathbf{p}}$ and $\bar{\mathbf{s}}$; explicitly[1]

$$\bar{\mathbf{p}} \triangleq [(\bar{\mathbf{p}}^0)^\mathsf{T}, \ldots, (\bar{\mathbf{p}}^{N_xN_y-1})^\mathsf{T}]^\mathsf{T} \in \mathbb{C}^{N_xN_yN_zN_s}$$
$$\bar{\mathbf{s}} \triangleq [(\bar{\mathbf{s}}^0)^\mathsf{T}, \ldots, (\bar{\mathbf{s}}^{N_xN_y-1})^\mathsf{T}]^\mathsf{T} \in \mathbb{C}^{N_xN_yN_kN_f}.$$

Now, we form the block-diagonal matrix Φ

$$\Phi \triangleq \operatorname{blkdiag}\left(\left\{\hat{\Phi}^l\right\}_{l=0}^{N_xN_y-1}\right) \in \mathbb{C}^{N_xN_yN_kN_F \times N_xN_yN_zN_s}.$$

Finally, we can write the vectorized form of the N-species forward model (5.7) as

$$\bar{\mathbf{s}} = \Phi\bar{\mathbf{p}}.$$

We call Φ the *N-species measurement matrix*. The block-diagonal structure of Φ illustrates the decomposition of range $\{\Phi\}$ into the direct sum of N_xN_y invariant subspaces,

$$\operatorname{range}\{\Phi\} = \operatorname{range}\left\{\Phi^1\right\} \oplus \ldots \oplus \operatorname{range}\left\{\Phi^{N_xN_y}\right\}, \tag{5.11}$$

where each subspace corresponds to one of the N_xN_y transverse Fourier frequencies $\mathbf{q}_\parallel$.

It exhibits a block structure that is similar to Φ^l. In a bit of overloaded notation, let $\hat{\mathbf{A}}_{n_F}$ be the

[1] Note that $\bar{\mathbf{p}}$ is not vec$(\mathbf{P})$, as vec$(\cdot)$ is defined with row-major ordering.

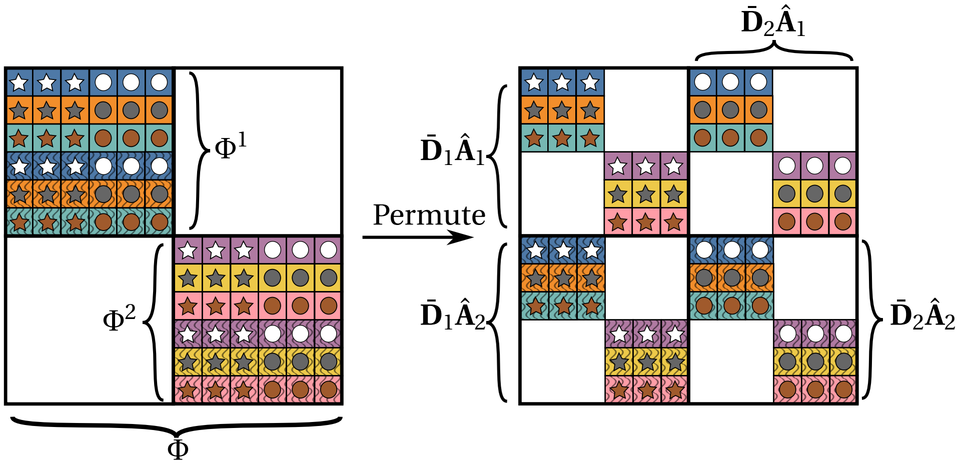

Φ^1
Φ^2
Φ
Permute
$\bar{\mathbf{D}}_2\hat{\mathbf{A}}_1$
$\bar{\mathbf{D}}_1\hat{\mathbf{A}}_1$
$\bar{\mathbf{D}}_1\hat{\mathbf{A}}_2$
$\bar{\mathbf{D}}_2\hat{\mathbf{A}}_2$

Figure 5.4: Row and column permutations bring the N-species measurement Φ into a block-matrix form similar to that of Φ^l. Here, color indicates the value of k_0 and $\mathbf{q}_\parallel$. Stars and circles denote species 1 and 2, respectively. Rows with solid (*resp.* wave-patterned) blocks correspond to measurements at the first (*resp.* second) focal plane. The blocks Φ^1 and Φ^2 are given by (5.9).

block-diagonal matrix $\hat{\mathbf{A}}_{n_F} = \text{blkdiag}\left(\{\hat{\mathbf{A}}^l_{n_F}\}_{l=0}^{N_xN_y-1}\right)$; see Fig. 5.2. With this definition, we have

$$\text{vec}\left(\sum_{n_z=0}^{N_z-1} \hat{\mathbf{A}}_{n_F}[:,:,n_z]\hat{\mathbf{p}}_{n_s}[:,n_z]\right) = \hat{\mathbf{A}}_{n_F}\text{vec}\left(\hat{\mathbf{P}}_{n_s}\right).$$

This is the discretized analogue of (3.3), with the added constraint that the target is non-dispersive. We call $\hat{\mathbf{A}}_{n_F}$ the *ISAM matrix*, as it models the action of ISAM on a discretized spatial density. We must also define $\bar{\mathbf{D}}_{n_s} = \mathbf{I}_{N_xN_Y} \otimes \mathbf{D}_{n_s}$, that has N_xN_y repeated copies of $\mathbf{h}_{n_s}$ along its diagonal; see Fig. 5.3. There exist permutation matrices Π_1, Π_2 such that

$$\Pi_1\Phi\Pi_2 = \begin{bmatrix} \bar{\mathbf{D}}_1\hat{\mathbf{A}}_1 & \bar{\mathbf{D}}_2\hat{\mathbf{A}}_1 & \dots & \bar{\mathbf{D}}_{N_s}\hat{\mathbf{A}}_1 \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\mathbf{D}}_1\hat{\mathbf{A}}_{N_F} & \bar{\mathbf{D}}_2\hat{\mathbf{A}}_{N_F} & \dots & \bar{\mathbf{D}}_{N_s}\hat{\mathbf{A}}_{N_F} \end{bmatrix}.$$

This relationship is illustrated in Fig. 5.4.

5.4.4 Construction using Khatri-Rao product

We briefly discuss an alternate construction of Φ^l that connects the N-species inverse problem to a broad range of related problems. We discuss these connections in Section 5.5.2.

Definition 5.2. *The* row-wise Khatri-Rao *product of matrices* $\mathbf{A} \in \mathbb{C}^{m \times n_1}$ *and* $\mathbf{B} \in \mathbb{C}^{m \times n_2}$ *is*

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} \mathbf{A}[1,:] \otimes \mathbf{B}[1,:] \\ \vdots \\ \mathbf{A}[m,:] \otimes \mathbf{B}[m,:] \end{bmatrix} \in \mathbb{C}^{m \times n_1 n_2},$$

i.e. each row of $\mathbf{A} \odot \mathbf{B}$ *is the Kronecker product of the corresponding rows of* $\mathbf{A}$ *and* $\mathbf{B}$.

We use the Khatri-Rao product to construct Φ^l. The first block-row of Φ^l is $\mathbf{H} \odot \hat{\mathbf{A}}_1^l$. To obtain all block-rows of Φ^l, we first stack the $\{\hat{\mathbf{A}}_{n_F}^l\}_{n_F=1}^{N_F}$ into the matrix $\bar{\mathbf{A}}^l \triangleq [(\hat{\mathbf{A}}_1^l)^\top \dots (\hat{\mathbf{A}}_{N_F}^l)^\top]^\top \in \mathbb{C}^{N_F N_k \times N_z}$. Next, stack N_F copies of $\mathbf{H}$ into $\bar{\mathbf{H}} \triangleq (\mathbb{1}_{N_F}^\top \otimes \mathbf{H}) = [\mathbf{H}^\top, \dots, \mathbf{H}^\top]^\top \in \mathbb{C}^{N_F N_k \times N_s}$. Now, $\Phi^l = \bar{\mathbf{H}} \odot \bar{\mathbf{A}}^l$. The complete matrix Φ can be constructed using row and column permutations.

5.5 The N-Species Inverse Problem

5.5.1 Preliminaries: The Single Species Case

Under the N-species model (5.7), the measurements at each focal plane are modeled as the sum of N_s independent ISAM experiments; thus, the ISAM matrices $\hat{\mathbf{A}}_{n_F}$ set fundamental limits on what can be imaged. Stated plainly, if a spatial density lies in the null space of $\hat{\mathbf{A}}_{n_F}$, then it will generate no measurement and thus cannot be imaged using the proposed method.

We have investigated the spectral properties of the continuous ISAM operator in Chapter 4. We are now interested in the properties of the discretized ISAM matrices.

Recall that the continuous formulation of Chapter 4 ignored the assumption that the susceptibility is compactly supported; however, the effects of compact support (or of a finite simulation volume) will be evident in the discretized setting.

A careful study of the spectral properties of $\hat{\mathbf{A}}_{n_F}^l$ is beyond the scope of this work. Instead, we combine a numerical study of these matrices with intuition obtained from the projection-slice interpretation of ISAM.

We computed the singular values of $\hat{\mathbf{A}}_{n_F}^l$ in the case of one transverse dimension, x, using the computational parameters listed in Table 5.1. The singular values are shown in Fig. 5.5, where k_x is determined from $q_x = \gamma^{-1}(l)$ using (5.6). While we do not form $\hat{\mathbf{A}}_{n_F}^l$ using the approximate kernel, the approximate kernel provides intuition for the behavior seen here. The largest singular values die off quickly as k_x increases, as expected due to the function $H\mathbf{k}_\parallel, k_0$ in (3.7). Moreover, for $|k_x| > 2k_b$, ISAM matrix is uniformly zero due to $\chi(\mathbf{k}_\parallel, 2k_0)$.

Table 5.1: Parameters for point target simulations.

N_x	192	L_x	423.6 μm	Δ_x	2.2 μm
N_z	384	L_z	282.4 μm	Δ_z	0.7 μm
N_k	384	k	$0.4\ \text{rad}\cdot\mu\text{m}^{-1}$	k_b	1.1
r_e	60	λ_{min}	$5.9\ \mu\text{m}^{-1}$	λ_{max}	$15.4\ \mu\text{m}^{-1}$
N_F	3	z_F	$[70, 140, 211]$ μm	NA	0.4

According to the approximate forward model, for $k_x = 0$ we obtain the (bandlimited) Fourier transform of the (compactly supported) weighted susceptibility. The eigenvalue spectrum of space-and-frequency limited Fourier operators has been studied, beginning with a series of papers by Slepian, Landau, and Pollak [139–141, 169, 170]. In the discrete case, the eigenvalue and singular value spectrum of space-and-frequency limited Discrete Fourier Transform (DFT) matrices have been studied; such matrices are submatrices formed by consecutive rows and columns of a DFT matrix [171–173]. The singular values of a space-and-frequency limited DFT matrix are divided into three distinct regions: (1) a region wherein the singular values are near one; (2) a transition region where the singular values decay exponentially; and (3) the remaining singular values are nearly zero. The number of singular values in the first region is called the *effective rank* and is written r_e. A direct application of Slepian-Pollak theory predicts [139, 173]

$$r_e = \frac{2(k_b - k_a)}{2\pi / L_z} = \frac{L_z}{\pi}(k_b - k_a). \tag{5.12}$$

For fixed $\mathbf{k}_\parallel$, the approximate ISAM operator can be viewed as a space-and-frequency limited Fourier operator with additional weighting in the spatial domain by $v(z)$ and in the frequency domain by $\vartheta\left(\mathbf{k}_\parallel, k_0\right)$. For each $\mathbf{k}_\parallel$ the operator is space-limited to a region of length L_z; this is due to assumption that η is compactly supported. Moreover, the operator is frequency-limited to the optical passband B. In the discretized setting, only $\mathbf{A}^0_{n_F}$ can be viewed as a (diagonally scaled) DFT matrix, as for $\mathbf{q}_\parallel \neq 0$ the resulting Fourier transform is not uniformly sampled.

We can use the theory of space-and-frequency limited DFT matrices to understand the behavior of the spectrum of $\hat{\mathbf{A}}^0_{n_F}$ as shown in Fig. 5.5. The singular values are broken into three regions: in the first region, the singular values decay exponentially, albeit at a rate slower than in the second region. The transition between the first and second regions still occurs at r_e. In the case of the parameters used in Fig. 5.5, we have $r_e = 60$, and the change in behavior at r_e is evident. The case of $k_x \neq 0$ is more complicated as the resulting Fourier transform is not uniformly sampled.

Recall that B is the set of observable Fourier components of the weighted susceptibility, $v(z -$

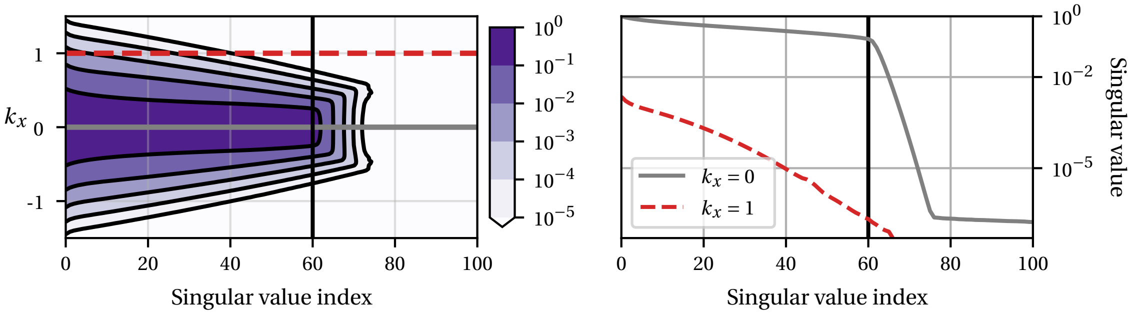


Figure 5.5: Left: Singular values of $\hat{\mathbf{A}}^l_{n_F}$. The coordinate k_x is obtained from $\gamma^{-1}(l)$ using (5.6). Right: singular values for $k_x = 0$ and $k_x = 1$. The vertical line marks the rank estimate (5.12). The focal plane is located at $z_F = 140$ µm. The remaining system parameters are listed in Table 5.1. The units of k_x are µrad)$^{-1}$.

$z_F)\,\hat{p}(\mathbf{k}_\parallel, z)$. A common practice in ISAM imaging is to ignore the axial weighting function and treat B as the observable Fourier components of the *unweighted* susceptibility (see, *e.g.* [100,101]). This is a reasonable approximation of the imaging system. To justify the approximation, note that $v(z)$ is strictly positive and slowly varying; thus the Fourier transform of the weighted and unweighted susceptibilities are roughly supported on the same set.

Using the same line of reasoning, we assume that null$\{\hat{\mathbf{A}}^l_{n_F}\}$ is invariant to the choice of focal plane z_F. This is reasonable when the focal planes are close to one another. Note that this is an implicit assumption in previous work on multi-focal ISAM [153].

5.5.2 Algebraic Conditions for a Unique Solution to (P1)

We return to the inverse problem under the N-species model, starting with (P1). In the discretized form of (P1), the $\mathbf{h}_{n_s}$ are fixed and known. The discretized forward model is $\bar{\mathbf{s}} = \Phi\bar{\mathbf{p}}$. As the $\mathbf{h}_{n_s}$ are fixed and known, the matrix Φ is completely determined, and recovery of $\bar{\mathbf{p}}$ is a linear inverse problem. Without additional constraints on the spatial densities, the existence and uniqueness of a solution is determined entirely by Φ. In this section, we establish algebraic conditions for existence and uniqueness of a solution in terms of the ISAM matrices, $\{\hat{\mathbf{A}}_{n_F}\}_{n_F=1}^{N_F}$, and the chemical spectra, $\{\mathbf{h}_{n_s}\}_{n_s=1}^{N_s}$. Earlier work on this problem claimed that $N_F \geq N_s$ and linear independence of the $\mathbf{h}_{n_s}$ is necessary and sufficient for unique recovery of the spatial densities $\hat{\mathbf{p}}_{n_s}$ within the optical passband [160]. While necessary, we show these two conditions are not sufficient.

We use the invariant subspace decomposition of Φ given by (5.11) to reduce the problem to the study of the "one-dimensional" problem $\bar{\mathbf{s}}^l = \Phi^l\bar{\mathbf{p}}^l$ for $l \in \{0, \ldots, N_xN_y - 1\}$, with Φ^l given by (5.9).

In what follows, the index l is fixed. We analyze the system independently for each transverse Fourier mode. The results can be applied block-by-block to pass to the full matrix Φ.

For each focal plane, the ISAM matrix $\hat{\mathbf{A}}^l_{n_F}$ is of size $N_k \times N_z$, where N_k is the number of wavenumber samples and N_z is the (axial) length of the discretized spatial density. Per Section 5.5.1, we assume the null space of $\hat{\mathbf{A}}^l_{n_F}$ is invariant to the choice of focal plane, thus for fixed l each matrix has the same rank. Let $r \triangleq \operatorname{rank}\{\hat{\mathbf{A}}^l_{n_F}\}$ for $n_F \in [N_F]$. We write the shared nullspace of the ISAM matrices as $\mathsf{N}^l \subseteq \mathbb{C}^{N_z}$; we have

$$\mathsf{N}^l \triangleq \operatorname{null}\left\{\hat{\mathbf{A}}^l_{n_F}\right\} \text{ for } n_F \in [N_F].$$

The optical passband is $\left(\mathsf{N}^l\right)^{\perp}$. Define the subspace

$$\bar{\mathsf{N}}^l \triangleq \mathsf{N}^l \times \mathsf{N}^l \ldots \times \mathsf{N}^l = \operatorname{span}\left\{ \bar{\mathbf{p}}^l = [(\hat{\mathbf{p}}^l_1)^\mathsf{T}, \ldots, (\hat{\mathbf{p}}^l_{N_s})^\mathsf{T}]^\mathsf{T} \,\middle|\, \hat{\mathbf{p}}^l_{n_s} \in \mathsf{N}^l, n_s \in [N_s]\right\} \subseteq \mathbb{C}^{N_s N_z}$$

of block vectors where each block is in N^l. The subspace $\left(\bar{\mathsf{N}}^l\right)^{\perp}$ consists of block vectors where each block lies in the optical passband, $\left(\mathsf{N}^l\right)^{\perp}$. In an abuse of notation, we refer to both $\left(\mathsf{N}^l\right)^{\perp}$ and $\left(\bar{\mathsf{N}}^l\right)^{\perp}$ as "the optical passband".

Using the N-species model, the measurements are a weighted sum of ISAM experiments; thus any objects that lie in $\bar{\mathsf{N}}^l$ will also be in $\operatorname{null}\{\Phi^l\}$. If an object cannot be imaged using ISAM, it cannot be imaged using Φ^l. We must consider uniqueness modulo $\bar{\mathsf{N}}^l$; our goal is to establish conditions such that these are the *only* objects that cannot be imaged using Φ^l. In this case, the N-species model does not introduce additional ambiguity and each species is correctly identified. We do no worse using the N-species model than if we were able to image the spatial densities independently using the ISAM system.

Let us pause to consider the geometry of a simple case: two species and a single focal plane. Here, $\Phi^l = [\mathbf{D}_1\hat{\mathbf{A}}^l_1, \mathbf{D}_2\hat{\mathbf{A}}^l_1]$ and $\bar{\mathbf{s}}^l = \Phi^l\bar{\mathbf{p}}^l = \mathbf{D}_1\hat{\mathbf{A}}^l_1\hat{\mathbf{p}}_1 + \mathbf{D}_2\hat{\mathbf{A}}^l_1\hat{\mathbf{p}}_2$. Clearly, if $\hat{\mathbf{p}}^l_1$ and $\hat{\mathbf{p}}^l_2$ are each in N^l, then $\bar{\mathbf{s}}^l = \mathbf{0}$. Suppose the $\mathbf{h}_{n_s}$ are non-zero for each index; then $\mathbf{D}_{n_s}$ is full rank. Using the formula for the rank of a partitioned matrix,

$$\begin{aligned}
\operatorname{rank}\left\{\Phi^l\right\} &= \operatorname{rank}\left\{[\mathbf{D}_1\hat{\mathbf{A}}^l_1, \mathbf{D}_2\hat{\mathbf{A}}^l_1]\right\} \\
&= \operatorname{rank}\left\{[\mathbf{D}_1\hat{\mathbf{A}}^l_1]\right\} + \operatorname{rank}\left\{[\mathbf{D}_2\hat{\mathbf{A}}^l_1]\right\} - \dim\left(\operatorname{range}\left\{\mathbf{D}_1\hat{\mathbf{A}}^l_1\right\} \cap \operatorname{range}\left\{\mathbf{D}_2\hat{\mathbf{A}}^l_1\right\}\right) \\
&= 2r - \dim\left(\operatorname{range}\left\{\mathbf{D}_1\hat{\mathbf{A}}^l_1\right\} \cap \operatorname{range}\left\{\mathbf{D}_2\hat{\mathbf{A}}^l_1\right\}\right).
\end{aligned}$$

The last term captures the interplay between the $\mathbf{D}_{n_s}$ and $\hat{\mathbf{A}}^l_1$. We want to find conditions under which this intersection is trivial. As we assume $\mathbf{D}_{n_s}$ is full rank, we can instead ask when $\operatorname{range}\{\hat{\mathbf{A}}^l_1\} \cap \operatorname{range}\{\mathbf{D}_1^{-1}\mathbf{D}_2\hat{\mathbf{A}}^l_1\}$ is trivial. Loosely speaking, when is multiplication by a diagonal matrix enough to perturb a subspace out of alignment with itself?

Next, we define our notion of uniqueness modulo the ISAM nullspace.

Definition 5.3. *The solution to* $\bar{\mathbf{s}}^l = \Phi^l \bar{\mathbf{p}}^l$ *is said to be* unique within the optical passband *if* $\Phi^l \mathbf{x} = \Phi^l \mathbf{y} \implies \mathbf{x} - \mathbf{y} \in \bar{\mathsf{N}}^l$. *Equivalently, there is a unique* $\bar{\mathbf{p}}^l \in \left(\bar{\mathsf{N}}^l\right)^\perp$ *such that* $\bar{\mathbf{s}}^l = \Phi^l \bar{\mathbf{p}}^l$.

This definition sets up an equivalence relation on the spatial densities: we treat two spatial densities as equivalent if their difference lies in $\bar{\mathsf{N}}^l$, the null space of the ISAM matrices. This is the component to which we are inherently are blind even in the single species case.

Next, we cast the problem into a form where we implicitly work in the optical passband $\left(\bar{\mathsf{N}}^l\right)^\perp$. Let $\mathbf{V}^l \in \mathbb{C}^{N_z \times r}$ be a basis for $\left(\mathsf{N}^l\right)^\perp$. We introduce a new set of matrices: the *restricted ISAM matrix* $\hat{\mathbf{B}}^l_{n_F} \triangleq \hat{\mathbf{A}}^l_{n_F} \mathbf{V}^l \in \mathbb{C}^{N_k \times r}$ is the restriction of $\hat{\mathbf{A}}^l_{n_F}$ to the subspace $\left(\mathsf{N}^l\right)^\perp$. Clearly, $\hat{\mathbf{B}}^l_{n_F}$ has full column rank. Similarly, $\mathbf{I}_{N_F} \otimes \mathbf{V}^l$ is a basis for $\left(\bar{\mathsf{N}}^l\right)^\perp$. We define the *restricted N-species matrix*

$$\tilde{\Phi}^l \triangleq \Phi^l (\mathbf{I}_{N_F} \otimes \mathbf{V}^l) \in \mathbb{C}^{N_F N_k \times N_s r}.$$

The question of unique recovery (within the optical passband) is determined entirely by this matrix, as stated in the following result.

Lemma 5.1. *Let* $\Phi^l \in \mathbb{C}^{N_F N_k \times N_s N_z}$ *and* $\operatorname{rank}\{\hat{\mathbf{A}}^l_{n_F}\} = r$ *for* $n_F \in N_F$. *The following statements are equivalent:*

(C1) *There is a unique* $\bar{\mathbf{p}}^l \in \left(\bar{\mathsf{N}}^l\right)^\perp$ *such that* $\bar{\mathbf{s}}^l = \Phi^l \bar{\mathbf{p}}^l$

(C2) $\operatorname{null}\left\{\Phi^l\right\} = \bar{\mathsf{N}}^l$

(C3) $\operatorname{rank}\left\{\tilde{\Phi}^l\right\} = N_s r$

Proof. See Appendix C.2. □

We can construct the restricted N-species matrix $\tilde{\Phi}^l$ using the Khatri-Rao product. Let $\bar{\mathbf{B}}^l \in \mathbb{C}^{N_F N_k \times r}$ be the matrix formed by stacking the restricted ISAM matrices $\hat{\mathbf{B}}^l_{n_F}$ into a single block column: $\bar{\mathbf{B}}^l \triangleq [(\hat{\mathbf{B}}^l_1)^\top, \ldots, (\hat{\mathbf{B}}^l_{N_F})^\top]^\top$. Recall $\bar{\mathbf{H}} = (\mathbb{1}^\top_{N_F} \otimes \mathbf{H}) = [\mathbf{H}^\top, \ldots, \mathbf{H}^\top] \in \mathbb{C}^{N_F N_k \times N_s}$. Now,

$$\tilde{\Phi}^l = \Phi^l (\mathbf{I}_{N_F} \otimes \mathbf{V}^l) = \begin{bmatrix} \mathbf{D}_1 \hat{\mathbf{A}}^l_1 \mathbf{V}^l & \ldots & \mathbf{D}_{N_s} \hat{\mathbf{A}}^l_1 \mathbf{V}^l \\ \vdots & \ddots & \vdots \\ \mathbf{D}_1 \hat{\mathbf{A}}^l_{N_F} \mathbf{V}^l & \ldots & \mathbf{D}_{N_s} \hat{\mathbf{A}}^l_{N_F} \mathbf{V}^l \end{bmatrix} = \begin{bmatrix} \mathbf{D}_1 \hat{\mathbf{B}}^l_1 & \ldots & \mathbf{D}_{N_s} \hat{\mathbf{B}}^l_1 \\ \vdots & \ddots & \vdots \\ \mathbf{D}_1 \hat{\mathbf{B}}^l_{N_F} & \ldots & \mathbf{D}_{N_s} \hat{\mathbf{B}}^l_{N_F} \end{bmatrix} = \bar{\mathbf{H}} \odot \bar{\mathbf{B}}^l, \qquad (5.13)$$

mirroring the construction of Φ^l in Section 5.4.4.

In what follows, we establish necessary and sufficient conditions for uniqueness within the optical passband.

Theorem 5.2 (Necessary Condition for Uniqueness)**.** *The solution to* $\bar{\mathbf{s}}^l = \Phi^l \bar{\mathbf{p}}^l$ *is unique within the optical passband only if*

(N1) $N_k N_F \geq N_s r$

(N2) The spectral profiles are linearly independent ($\operatorname{rank}\{\mathbf{H}\} = N_s$*)*

(N3) No row of $\bar{\mathbf{B}}^l$ *is orthogonal to all remaining rows*

(N4) For every subset $J \subset [N_k]$ *with* $N_s \leq |J| < N_s r / N_F$ *and* $\operatorname{rank}\{\mathbf{H}^J\} = N_s$*, we have* $\operatorname{rank}\{\mathbf{H}^{J^c}\} \geq N_s - \frac{N_F}{r}|J|$

(N5) $\sum_{i=1}^{N_k} \operatorname{rank}\left\{\left[\hat{\mathbf{B}}_1^l[i,:]^\top, \ldots, \hat{\mathbf{B}}_{N_F}^l[i,:]^\top\right]^\top\right\} \geq N_s r$

Proof. See Appendix C.3. □

Let us pause to interpret these conditions.

In the single-species case, (N1) reduces to $N_k \geq r$; *i.e.* we must measure enough wavenumbers such that the single-species ISAM problem is well-posed. Interestingly, (N1) does not require that $N_F \geq N_s$: recovery of N_s species is possible from a single focal plane, provided the measurements are oversampled in wavenumber. This behavior can be seen in the numerical experiments described in Section 5.5.3

Condition (N2) is unsurprising. If the spectral profiles are linearly dependent, the N-species representation of a susceptibility is not unique and the spatial densities cannot be uniquely determined.

Condition (N3) is less transparent, but can be argued to hold by the underlying physics. If (N3) is violated, there must be an object that scatters at only one of the measured wavenumbers and is non-scattering for the rest. In the continuous setting, scattered fields are analytic functions of k_0; thus if an object is non-scattering over an interval of wavenumbers, it must be non-scattering for all k_0 [103, 174]. In the discretized setting we lose the analytic properties of scattered waves. In our experience, however, condition (N3) holds.

Condition (N4) requires the spectral profiles to be sufficiently diverse: linear independence is not enough. As an example, consider $N_s = 2, N_F = 1$, and take $\mathbf{h}_1 = [1, 1, \ldots, 1]^\top$ and $\mathbf{h}_2 = [2, 1, \ldots, 1]^\top$. These spectra are linearly independent, but $\mathbf{D}_1 \hat{\mathbf{A}}_1^l$ and $\mathbf{D}_2 \hat{\mathbf{A}}_1^l$ differ by only one row; thus $\operatorname{rank}\{\tilde{\Phi}^l\} \leq r + 1$, failing (C3) of Lemma 5.1. Spectral diversity is necessary to push $\operatorname{range}\{\mathbf{D}_1 \hat{\mathbf{A}}_1^l\}$ out of alignment with $\operatorname{range}\{\mathbf{D}_2 \hat{\mathbf{A}}_1^l\}$. "Good" spectral profiles are not too concentrated on any small set of indices.

The final condition, (N5), is a requirement on the diversity of measurements comprising the restricted ISAM matrices. When $N_k N_F = N_s r$, (N5) requires that the collection of measurement

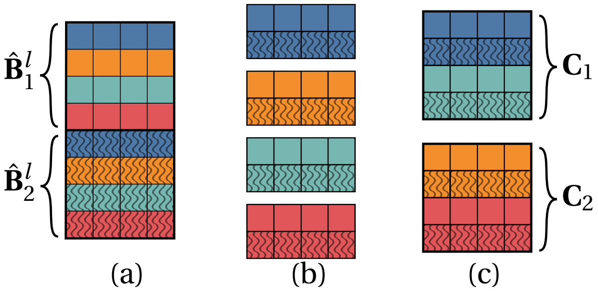


Figure 5.6: Comparing (N5) and Theorem 5.3 for $N_F = 2$ and $r = 4$. Color indicates the value of k_0. (a) The matrix $\bar{\mathbf{B}}^l$. Rows with solid (*resp.* wave-patterned) blocks correspond to measurements at the first (*resp.* second) focal plane. (b) Condition (N5) requires that the sum of the ranks of each 2×4 block of the same color must be at least $4N_s$. (c) A possible partitioning of the rows of $\bar{\mathbf{B}}^l$ as described in Theorem 5.3. If both $\mathbf{C}_1$ and $\mathbf{C}_2$, as defined in (5.14), have full rank for generic chemical species the solution to $\bar{\mathbf{s}}^l = \Phi^l \bar{\mathbf{p}}^l$ is unique within the optical passband with probability one.

vectors corresponding to a given wavenumber be linearly independent: each new focal plane must provide new and informative measurements. This partitioning is illustrated in Fig. 5.6.

Next, we establish a sufficient condition for unique recovery within the optical passband. First, we note that no conditions on $\bar{\mathbf{B}}^l$ or $\mathbf{H}$ independently are sufficient to ensure there is a unique solution within the optical passband. Consider again the two-species, one focal plane case: $\tilde{\Phi}^l = [\mathbf{D}_1\hat{\mathbf{B}}_1^l, \mathbf{D}_2\hat{\mathbf{B}}_1^l]$, with $\mathbf{D}_i = \operatorname{diag}(\mathbf{h}_i)$. Suppose $\mathbf{h}_1$ is fixed and choose vectors $\mathbf{w}, \mathbf{v} \in \mathbb{C}^r$ such that no entry of $\hat{\mathbf{B}}_1^l \mathbf{v}$ is zero. Set $\mathbf{h}_2 = (\mathbf{D}_1\hat{\mathbf{B}}_1^l\mathbf{w}) / (\hat{\mathbf{B}}_1^l\mathbf{v})$ where the division is taken elementwise. With this construction, $\mathbf{D}_2\hat{\mathbf{B}}_1^l\mathbf{v} = \mathbf{D}_1\hat{\mathbf{B}}_1^l\mathbf{w}$, and thus $\operatorname{rank}\{\tilde{\Phi}^l\} \leq 2r - 1$, failing (C3) of Lemma 5.1.

These spectral profiles were carefully chosen to make $\tilde{\Phi}^l$ lose rank. Fortunately, we are unlikely to encounter such objects in practice. The following definition makes this argument precise.

Definition 5.4. *A property that depends on the spectral profiles* $\mathbf{H} \in \mathbb{C}^{N_k \times N_s}$ *is said to hold* generically*, or* for generic $\mathbf{H}$*, if the set for which it fails to hold has Lebesgue measure zero and is nowhere dense in* $\mathbb{C}^{N_k \times N_s}$.

If a property that holds generically, it holds with probability one if the spectral profiles are drawn independently from a distribution that is absolutely continuous with respect to the Lebesgue measure in $\mathbb{C}^{N_k \times N_s}$; for instance, when the entries of $\mathbf{H}$ are drawn i.i.d. from the Gaussian distribution. Moreover, the property exhibits a degree of robustness: if it holds for a particular $\mathbf{H}'$, then it holds in an open ball around $\mathbf{H}'$ and will continue to hold given sufficiently small perturbations to $\mathbf{H}'$.

Theorem 5.3 (Sufficient Condition for Uniqueness)**.** *Suppose $N_k \geq r$ and $N_F \geq N_s$. Let $r = N_F m$ for an integer m. If there exists a collection $\{J_i \subset [N_k]\}_{i=1}^{N_F}$ of disjoint sets, each of cardinality*

$|J_i| = r/N_F$, *such that*

$$\mathbf{C}_i \triangleq \begin{bmatrix} \hat{\mathbf{B}}_1^l[J_i, :\,] \\ \vdots \\ \hat{\mathbf{B}}_{N_F}^l[J_i, :\,] \end{bmatrix} \in \mathbb{C}^{r \times r} \tag{5.14}$$

is full rank for each $i \in [N_F]$, *then for generic* $\mathbf{H}$ *the solution to* $\bar{\mathbf{s}}^l = \Phi^l \bar{\mathbf{p}}^l$ *is unique within the optical passband.*

Proof. See Appendix C.4. □

An illustration of the matrices $\mathbf{C}_i$ is shown in Fig. 5.6(c). Note that the necessary condition (N5) coincides with the sufficient condition of Theorem 5.3 in the case of $N_k = N_F = r = N_s$, which is the limit of scanning confocal spectroscopic acquisition discussed in Section 5.1.

Theorem 5.3 can be stated in terms of a more familiar, but more restrictive, property on $\bar{\mathbf{B}}^l$.

Definition 5.5. *The* Kruskal (row) rank *of a matrix* $\mathbf{X} \in \mathbb{C}^{n \times m}$, *written* $\operatorname{krank}\{\mathbf{X}\}$, *is the largest* k *such that every set of* k *rows of* $\mathbf{X}$ *are linearly independent. The matrix* $\mathbf{X}$ *is said to have* full Kruskal rank *if* $\operatorname{krank}\{\mathbf{X}\} = \max\{n, m\}$.

Corollary 5.4. *If* $\bar{\mathbf{B}}^l \in \mathbb{C}^{N_k N_F \times r}$ *has full Kruskal rank, then for generic* $\mathbf{H}$ *the solution to* $\bar{\mathbf{s}}^l = \Phi^l \bar{\mathbf{p}}^l$ *is unique within the optical passband.*

The Khatri-Rao structure of Φ provides a link between the N-species inverse problem and topics in tensor factorization, communications, and sensor networks, among others [175–181]. For example, the rank and Kruskal rank of the Khatri-Rao product has implications for the uniqueness of certain tensor factorizations. Properties of the Khatri-Rao product are an active area of research. For generic matrices $\mathbf{X}$ and $\mathbf{Y}$, it is known that $\operatorname{krank}\{\mathbf{X} \odot \mathbf{Y}\} = \operatorname{krank}\{\mathbf{X}\} \operatorname{krank}\{\mathbf{Y}\}$. Bhaskara *et al.* provide bounds on the smallest singular value of the Khatri-Rao product of random matrices [180]. Recent work has investigated the restricted isometry property of the Khatri-Rao product of random matrices [176–178].

These results do not directly apply to our problem. We are interested in properties of $\tilde{\Phi}^l = \bar{\mathbf{H}} \odot \bar{\mathbf{B}}^l$. As $\bar{\mathbf{B}}^l$ is determined by the physics and imaging geometry, we cannot choose this matrix generically or randomly. Even $\bar{\mathbf{H}}$ cannot be chosen generically, as $\bar{\mathbf{H}} = (\mathbb{1}_{N_F}^\top \otimes \mathbf{H})$; only the matrix $\mathbf{H}$ can be chosen generically. Translating new results on the Khatri-Rao product to our setting remains a topic for further investigation.

5.5.3 Stability And Conditioning of (P1)

The results of Section 5.5.2 tell us that the solution to $\bar{\mathbf{s}}^l = \Phi^l \bar{\mathbf{p}}^l$ is almost always unique, but say little about the stability of the problem. We must always deal with "noisy" measurements– not just instrumentation noise, but also "noise" due to modeling error, *e.g.* multiple scattering and spatial-spectral coupling not captured by the N-species model.

In this section, we numerically investigate the behavior of the singular values of the N-species matrix Φ for the case three-species case ($N_s = 3$) in two spatial dimensions. We use the computational parameters listed in Table 5.1, except for NA and N_F, which vary. The singular values of the ISAM matrix formed using these computational parameters were investigated in Section 5.5.1 and plotted in Fig. 5.5.

We computed the singular values of each block-matrix Φ^l (5.9) and plot the results in Fig. 5.7. Recall as a function of k_x is determined from $q_x = \gamma^{-1}(l)$ using (5.6). As expected, higher transverse spatial frequencies are present as NA increases. Only the first $N_F r_e$ singular values are appreciable. The low-frequency components achieve rank $3r_e$ for $N_F = 3$, and adding focal planes improves the conditioning of Φ. Note that even in the case of a single focal plane, the $3r_e$-th singular value of Φ^0 is non-zero; as previously discussed, $N_F \geq N_s$ is not necessary for a unique solution.

We investigated the singular values of the block corresponding to $k_x = 0$ for a variety of chemical species and a varying number of focal planes. We used a library of 20 experimentally acquired chemical spectra[2] provided through the IARPA SILMARILS project. We randomly selected three species from the library, formed Φ^0, and computed the singular values of this matrix. We scaled Φ^0 to have unit spectral norm. This procedure was repeated for 200 realizations. The resulting singular values are plotted in Fig. 5.8; the borders of the shaded region are the best and worst realizations for each choice of N_F.

We repeated the same procedure using random spectral profiles. The real part of the spectral profile was drawn i.i.d. from the standard normal distribution and the imaginary part was chosen uniformly over $[0,1]$. The results are plotted in Fig. 5.8. Clearly, these un-physical spectra lead to better conditioned Φ^0, and there is little difference in the best and worst realizations. Study of the system using random spectral profiles may lead to a useful upper bound on system performance.

[2] These include caffeine, acetaminophen, warfarin, monosodium glutamate (MSG), sucrose, naproxen, potassium chlorate, polyvinylidene fluoride (PVDF), aspartame, lactose, melatonin, ethylenediaminetetraacetic acid (EDTA), creatine, diazepam, biotin, fructose, pectin, glycine, beta carotene, hydroxypropyl cellulose.

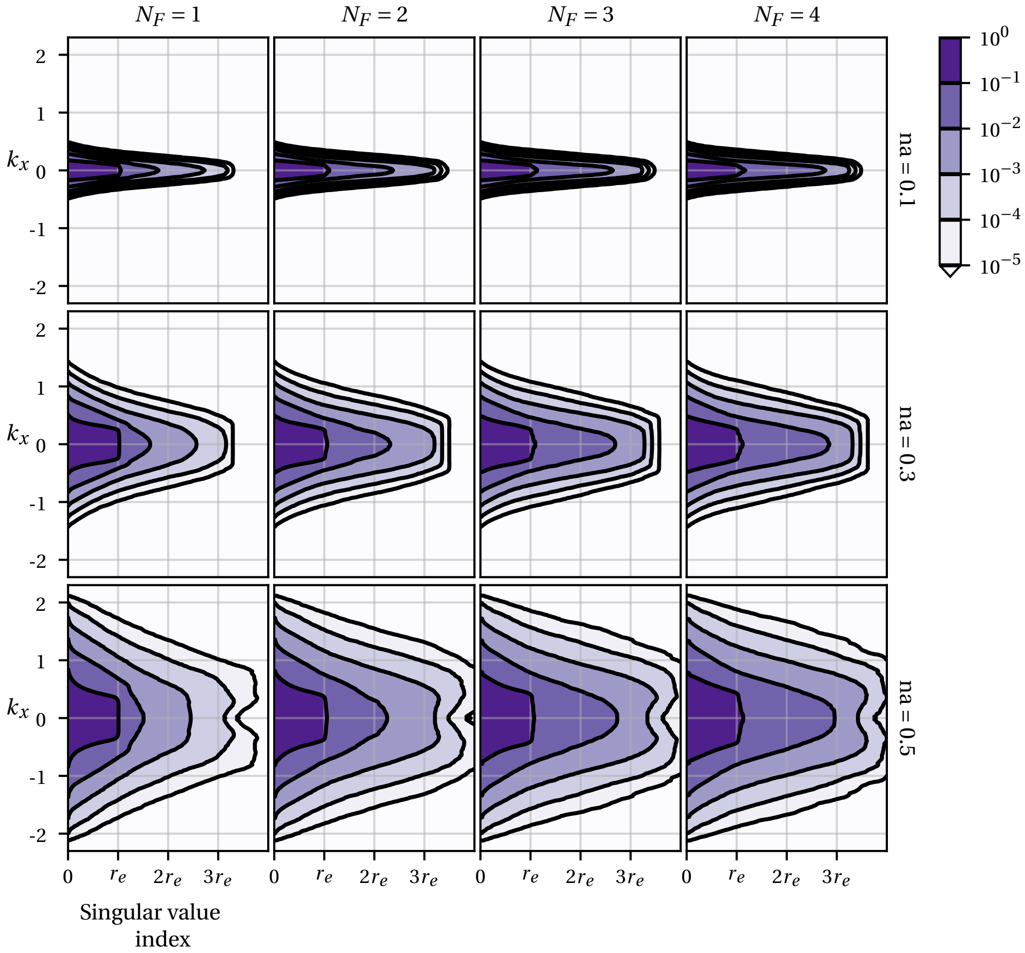


Figure 5.7: Singular values of Φ^l. The coordinate k_x is obtained from $\gamma^{-1}(l)$ using (5.6). Three species are present: caffeine, acetaminophen, and warfarin. System parameters listed in Table 5.1.

5.5.4 Algebraic Conditions for (P2)

We now focus on the case (P2), wherein the target comprises N_s chemical species drawn from a "dictionary" of $M_s > N_s$ possible spectra. This problem can be viewed as an instance of (P1), in which case Theorem 5.2 requires that number of focal planes is chosen such that $N_F N_k \geq M_s r$. This is undesirable if M_s is much larger than N_s. This approach ignores the constraint that only N_s chemicals are present in the sample; by incorporating this side information, we relax our condition on N_F. This structure is known as *block sparsity.*

Definition 5.6. *The block vector* $\bar{\mathbf{p}}^l = [\hat{\mathbf{p}}_1^\top, \ldots, \hat{\mathbf{p}}_2^\top, \hat{\mathbf{p}}_{M_s}^\top]^\top$ *is said to be* block-K sparse *if the set* $\{i : \|\hat{\mathbf{p}}_i\|_2 > 0\}$ *has cardinality at most* K*.*

Block sparsity is a natural fit for our problem; we define the n_s-th block to be the n_s-th spatial

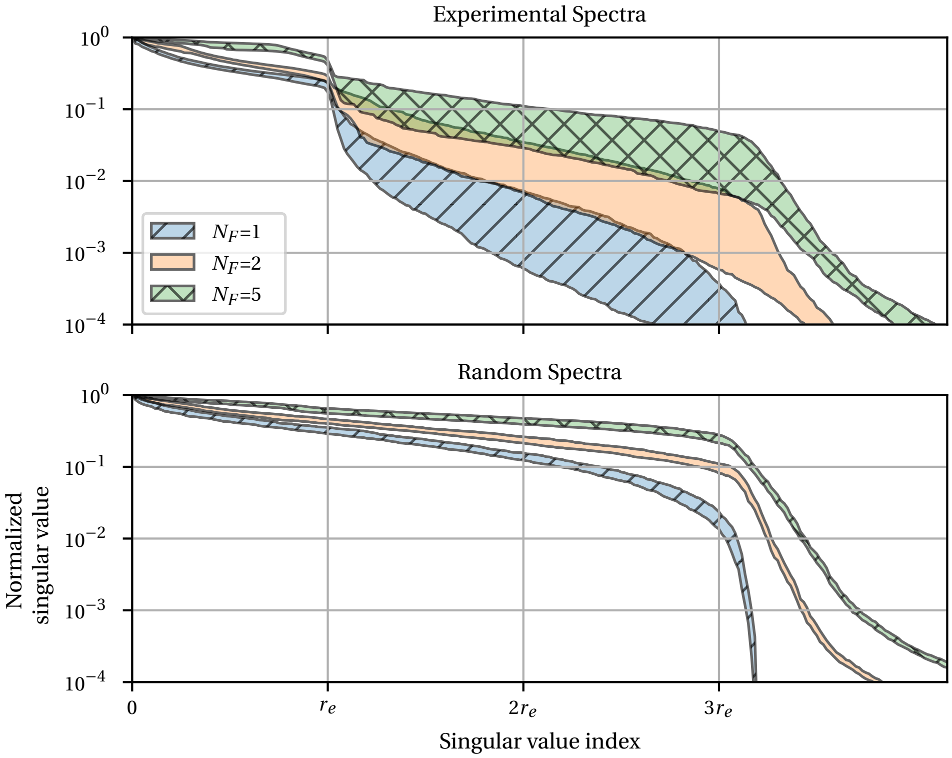


Figure 5.8: Singular values of Φ^0 for various combinations of chemical species. The shaded area lies in between the best and worst realizations. System parameters listed in Table 5.1. Top: singular values using experimentally acquired spectral profiles. Bottom: singular values using random Gaussian spectral profiles.

density $\hat{\mathbf{p}}_{n_s}$, corresponding to the n_s-th species in the dictionary. Note that block sparsity does not require the blocks themselves (*i.e.*, the $\{\hat{\mathbf{p}}_{n_s}\}_{n_s=1}^{N_s}$) to be sparse.

Conditions for unique recovery of block-sparse vectors have been studied [182–185]. Eldar and Mishali [184] developed a straightforward condition for unique recovery that suits our needs:

Lemma 5.5. *[184, Proposition 1] There is a unique block-N_s sparse solution to $\bar{\mathbf{s}}^l = \Phi^l \bar{\mathbf{p}}^l$ if and only if $\Phi^l \mathbf{v} \neq 0$ for any non-zero $\mathbf{v}$ that is block-$2N_s$ sparse.*

We can easily translate Lemma 5.5 into our setting.

Theorem 5.6 (Unique recovery with block-sparsity)**.** *For generic $\mathbf{H}$, there is a unique block-N_s sparse vector $\bar{\mathbf{p}}^l$ consistent with measurements $\bar{\mathbf{s}}^l = \Phi^l \bar{\mathbf{p}}^l$ if $N_F \geq 2N_s$ and $\bar{\mathbf{B}}^l$ contains $2N_s$ disjoint sets of* rank$\{\bar{\mathbf{B}}^l\}$ *linearly independent rows.*

Proof. Take $N_F \geq 2N_s$. Let $\mathbf{v}$ be a block-$2N_s$ sparse vector with $\Gamma = [\gamma_1, \ldots, \gamma_{2N_s}]^\top$ indexing the $2N_s$ non-zero blocks of $\mathbf{v}$. Let $\Phi_\Gamma \in \mathbb{C}^{N_F N_k \times 2N_s N_z}$ be the restriction of Φ to the $2N_s$ columns indexed by Γ, and let $\mathbf{v}_\Gamma \in \mathbb{C}^{2N_sN_z}$ be the non-zero elements of $\mathbf{v}$. By Theorem 5.3, Φ_Γ is generically full column rank; and generically $\Phi\mathbf{v} = \Phi_\Gamma\mathbf{v}_\Gamma \neq 0$. Applying Lemma 5.5 completes the proof. □

5.5.5 Computational Recovery

In the single-species case, the approximate form of the ISAM operator allows for the use of the non-iterative Fourier inversion algorithm Algorithm 3, as described in Section 3.6. This does not carry over to the multi-species case.

We recover the collection of spatial densities $\hat{\mathbf{P}}$ by solving the penalized least squares problem

$$\arg\min_{\hat{\mathbf{P}}} \frac{1}{2}\|\bar{\mathbf{s}} - \Phi\bar{\mathbf{p}}\|_2^2 + \lambda_r R(\hat{\mathbf{P}}). \tag{5.15}$$

The first term is known as the *data fidelity* term. It ensures the observed data $\bar{\mathbf{s}}$ and "re-imaged" solution $\Phi\bar{\mathbf{p}}$ are consistent. More sophisticated data fidelity terms can be used to model the effects of shot noise, background signal, and more [123], but these are beyond the scope of this work.

The functional $R : \mathbb{C}^{N_x \times N_y \times N_z \times N_s} \to \mathbb{R}$ regularizes the inverse problem and encodes any constraints or *a priori* assumptions regarding the spatial densities. Tikhonov regularization corresponds to $R(\hat{\mathbf{P}}) = \sum_{n_s=1}^{N_s} \|\hat{\mathbf{p}}_{n_s}\|_2^2$. Alternatively, solutions that are sparse in a transform domain are obtained by setting $R(\hat{\mathbf{P}}) = \sum_{n_s=1}^{N_s} \|\mathbf{C}\hat{\mathbf{p}}_{n_s}\|_1$, where $\mathbf{C}$ is a sparsifying transform, *e.g.* a wavelet transform. Finally, the mixed ℓ_1/ℓ_2 norm $\sum_{n_s=1}^{N_s} \|\hat{\mathbf{p}}_{n_s}\|_2$ encourages solutions that are block-sparse; that is, solutions with a minimal number of active species. The non-negative scalar λ_r balances the influence of the data fidelity and regularization terms.

The method used to solve (5.15) depends on the chosen regularizer. In the case of Tikhonov regularization, (5.15) reduces to the solution of the linear system

$$(\Phi^{\mathsf{H}}\Phi + \lambda_r\mathbf{I})\bar{\mathbf{p}} = \Phi^{\mathsf{H}}\bar{\mathbf{s}}, \tag{5.16}$$

where $\mathbf{I}$ is the $N_F N_k N_x N_y \times N_F N_k N_x N_y$ identity matrix. The matrix $\Phi^{\mathsf{H}}\Phi \in \mathbb{C}^{N_F N_k N_x N_y \times N_F N_k N_x N_y}$ is too large to store, much less invert, an iterative solution is required. The conjugate gradient (CG) algorithm works well in practice. CG requires only matrix-vector products with Φ and Φ^{H}. These matrices are not explicitly formed; only the coefficients $\hat{\mathbf{A}}_{n_F}[\mathbf{q}_\parallel, n_k, n_z]$ are precomputed and stored. Similarly, the $N_x N_y N_k \times N_x N_y N_k$ matrices $\bar{\mathbf{D}}_{n_s}$ are not formed; only the spectral profiles are stored, and products with $\bar{\mathbf{D}}_{n_s}$ are computed by elementwise multiplication. We compute the matrix-vector products with Φ in a block-wise fashion. The vector $\bar{\mathbf{y}} = \Phi\bar{\mathbf{p}}$ consists of $N_x N_y$ blocks $\bar{\mathbf{y}}^l_{n_F}$, where $l = \gamma(\mathbf{q}_\parallel) \in \{0, \ldots, N_x N_y - 1\}$, $n_F \in [N_F]$, and

$$\hat{\mathbf{y}}^l_{n_F} = \sum_{n_s=1}^{N_s} \mathbf{D}_{n_s} \hat{\mathbf{A}}^l_{n_F} \hat{\mathbf{p}}^l_{n_s}.$$

Assuming the spatial densities are already in the transverse Fourier domain, computing products with the N-species matrix $\Phi \in \mathbb{C}^{N_F N_k N_x N_y \times N_s N_z N_x N_y}$ in this way requires $O(N_x N_y N_z N_F N_k N_s)$ FLOPS, rather than $O(N_x^2 N_y^2 N_z N_F N_k N_s)$ FLOPS required if we ignore the block structure in Φ. Similarly, $\bar{\mathbf{w}} = \Phi^{\mathsf{H}} \bar{\mathbf{y}}$ consists of blocks $\hat{\mathbf{w}}^l_{n_s}$ with $n_s \in N_s$, where the block is computed as

$$\hat{\mathbf{w}}^l_{n_s} = \sum_{n_F=1}^{N_F} (\hat{\mathbf{A}}^l_{n_F})^{\mathsf{H}} \mathbf{D}^{\mathsf{H}}_{n_s} \hat{\mathbf{y}}^l_{n_F}.$$

Many sparsity-promoting regularizers are non-differentiable. In this case, proximal methods such as FISTA [89] or the Alternating Direction Method of Multipliers (ADMM) [126–128] are attractive. This class of algorithms decomposes the problem (5.15) into a sequence of simpler subproblems. The solution of a linear system similar to (5.16) is often a key ingredient of such algorithms.

5.6 Simulations

We now describe two simulations used to validate the proposed approach. For simplicity, we consider only two spatial dimensions: one transverse (x) and one axial (z).

Preliminary work on the N-species model suffers from three unrealistic assumptions [160]. The simulations used unrealistic wavelength ranges, leading to nearly complete coverage of Fourier space. This removes the large null space present in $\mathbf{A}_{n_F}$ and simplifies the reconstruction problem. Secondly, the phantoms used satisfied the N-species model exactly; no spectral noise was considered. Finally, the synthetic data used in the simulations was generated data using the asymptotic approximation to the ISAM operator (3.6), and thus under the first Born approximation. This neglects multiple scattering, absorption, and the discrepancy between the exact and approximate ISAM models. As a consequence, the simulations present an overly optimistic view of the proposed imaging modality.

We generate synthetic data using accurate physical models and system parameters. Our synthetic data includes multiple scattering and absorption effects—only the inversion is performed under the Born approximation. Further, our simulated targets do not precisely follow the N-species model; instead, there are position-dependent spectral variations within each

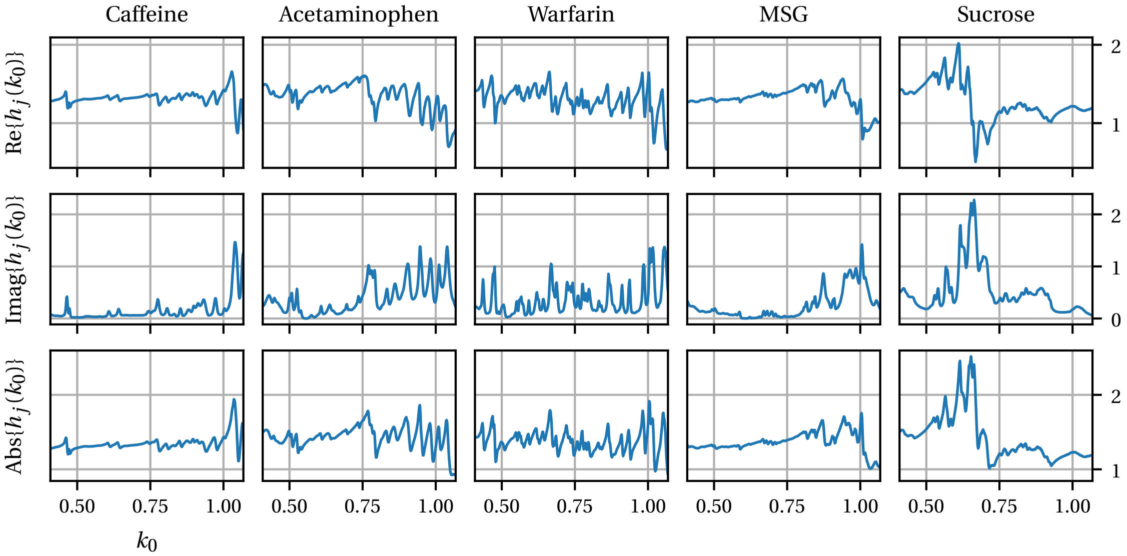


Figure 5.9: Spectral profiles for the five chemicals used in point scattering simulations.

species. In particular, we simulate an object of the form $\eta(\mathbf{r}, k_0) = \sum_{n_s=1}^{N_s} p_{n_s}(\mathbf{r}) h_{n_s}(\mathbf{r}, k_0)$, where $h_{n_s}(\mathbf{r}, k_0) = h_{n_s}(k_0) + e_{n_s}(\mathbf{r}, k_0)$ and $e_{n_s}(\mathbf{r}, k_0) \sim \mathcal{CN}(0, \xi_{n_s})$ is a circular complex Gaussian random variable [186].

The minimization problem (5.15) is solved on an NVidia Titan X GPU using a combination of Python and CUDA [87, 88].

5.6.1 Point Targets

We formed a spectral library of five chemicals using refractive index data provided through the IARPA SILMARILS project. The corresponding spectral profiles are plotted in Fig. 5.9. The target consisted of 50 point scatterers. Each point scatterer is associated to one chemical species; only three species (out of the five possible) are present. We do not know *a priori* which chemicals are present.

We generated measurements using the Foldy-Lax model, which includes multiple scattering effects [109]. Data was generated at three focal planes in a 420 × 280 µm volume according to the parameters in Table 5.1. The source power spectrum was flat over $[k_a, k_b]$. This combination of parameters—three active species, three focal planes, and a library of five possible species—corresponds to the case of (P2).

To assess the deviation from the single scattering model, we generated two sets of measurements using the same target. The first set of measurements, denoted $\mathbf{s}$, uses the Foldy-Lax method and incorporates multiple scattering. The second, $\mathbf{s}_B$, is generated using the Born ap-

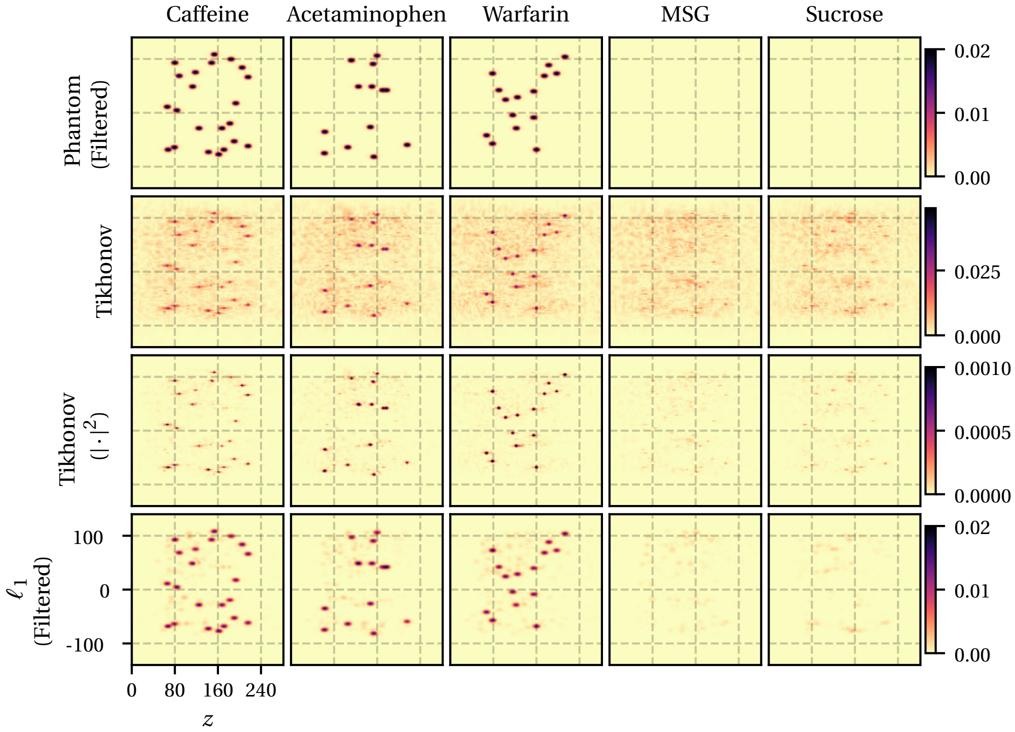


Figure 5.10: Reconstructions of point scatterers described in Section 5.6.1.

proximation and thus includes only single scattering events. The ratio $\|\mathbf{s}-\mathbf{s}_B\|_2/\|\mathbf{s}_B\|_2$ indicates that more than 20% of the energy in $\mathbf{s}$ comes from multiple scattering events.

We performed two sets of simulations: the first using Tikhonov regularization and the second using sparsity-promoting regularization. In the latter case, motivated by the spatial-domain sparsity of the target, we set $R(\mathbf{P}) = \sum_{n_s=1}^{5}\|\mathbf{p}_{n_s}\|_1$. In the Tikhonov case, we performed 300 iterations of conjugate gradient on the normal equations with $\lambda_r = 10^{-5}$. In the case of ℓ_1 regularization, we used 2000 iterations of the FISTA algorithm with $\lambda_r = 10^{-3}$. Both cases terminated in under one minute.

The magnitude of the reconstructed spatial densities are shown in Fig. 5.10. Recall that the surface of observable Fourier components is restricted to $k_z < 0$. As such, any linear reconstruction method (*e.g.*, Tikhonov-regularized least squares) will produce a complex-valued image; we display only the magnitude and squared magnitude of the recovered signal. For visualization purposes we have projected the point-target phantom onto the optical passband. In both cases, the reconstructed targets are correctly spatially localized and identified with the correct species.

The Tikhonov regularized reconstruction consists of the point scatterers sitting on top of a "noisy" background. The background is primarily due to multiple scattering effects and spectral variations which are not captured by our forward model. This background term is

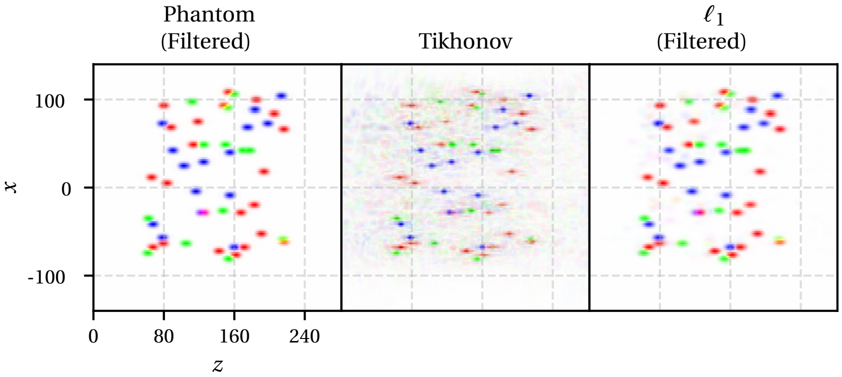


Figure 5.11: Visualizing the reconstructed point targets by assigning the three active species to a RGB channel. Red: caffeine, Green: acetaminophen, Blue: warfarin. The two species that are not present (MSG, Sucrose) were ignored.

distributed across all five possible species; however, the recovered point scatterers are associated to the correct species. The background is eliminated when viewing the squared modulus of the reconstruction.

The ℓ_1 regularized reconstruction suppresses the background term. There is nearly perfect agreement between the true target and the reconstructed target, despite taking data at only three, rather than five, focal planes. The sparsity of the target, coupled with the ℓ_1 regularization, successfully eliminates artifacts due to multiple scattering.

For visualization purposes we map the three active species to the red, green, and blue channels of an RGB image. The filtered phantom, Tikhonov, and filtered ℓ_1 reconstructions are shown in Fig. 5.11.

5.6.2 Cell Phantom

Next, we evaluated the ability to image extended targets. Our target is the cellular phantom shown in Fig. 5.12, which comprises three chemical species. Our spectral library contains five total species.

We generated synthetic measurements by solving the Lipmann-Schwinger equation (see, *e.g.*, [109]) using the using the Multi-Level Fast Multipole Algorithm (MLFMA) [187]. The data is not generated under the Born approximation, and thus includes multiple scattering and absorption phenomenon not captured using our forward model. We use a version of the MLFMA specialized for simulating two spatial dimensions[3] [188, 189].

[3] We thank Mert Hidayetoglu for providing the MLFMA code.

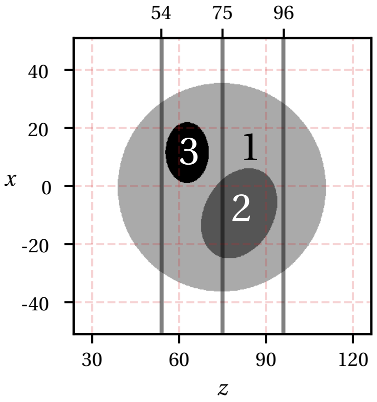


Figure 5.12: Three-species cell phantom. The detector plane is plane located at $z = 0$. Solid vertical lines denote the three focal planes. All units are μm.

Table 5.2: Parameters for cell phantom simulation.

N_x	256	L_x	150.0 μm	Δ_x	0.6 μm
N_z	256	L_z	150.0 μm	Δ_z	0.6 μm
N_k	256	k_a	$0.7\ \text{rad}\cdot\mu\text{m}^{-1}$	k_b	$2.1\ \text{rad}\cdot\mu\text{m}^{-1}$
r_e	67	$\lambda_{\min}$	$3.0\ \mu\text{m}^{-1}$	$\lambda_{\max}$	$9.0\ \mu\text{m}^{-1}$
N_F	3	z_F	[54, 75, 96] μm	NA	0.5

We generated measurements for only three focal planes; the relevant computational parameters are listed in Table 5.2. We used synthetic spectral profiles: each $\mathbf{h}_{n_s}$ is generated according to (5.4) where $N_l = 100$ and the remaining parameters are randomly chosen. In particular, $\sigma_{i,n} \sim \text{Unif}[0, 0.1]$, $k_{i,n} \sim \text{Unif}[1.2\pi, 4.4\pi]$, $\gamma_{i,n} \sim \text{Unif}[2\pi\times 10^{-3}, 4\pi\times 10^{-2}]$, where $\text{Unif}[a, b]$ is the uniform distribution over the interval $[a, b]$. The spectral profiles are plotted in Fig. 5.13.

The first-order Born approximation is valid only if the total phase change between the incident field and the field inside the sample is less than π—this implies that the object should be either weakly scattering or small in spatial extent [190, 191]. The proposed phantom is neither. To investigate the effect on scattering strength on the reconstructed images, we generated synthetic measurements for the scaled object $\delta\eta(\mathbf{r}^{(o)}, k_0)$ where $0 < \delta \leq 1$. By reducing δ, we reduce the scattering strength and eventually fall into a regime where the first-order Born approximation holds.

We used Tikhonov regularization with $\lambda_r = 1\times 10^{-4}$ and 500 iterations of the conjugate-gradient algorithm. The resulting reconstructions are shown in Fig. 5.14. The top row illustrates the projection of the phantom onto the optical passband; this serves as the “gold standard” for our Tikhonov-regularized reconstructions. The remaining rows are the reconstructed images. As

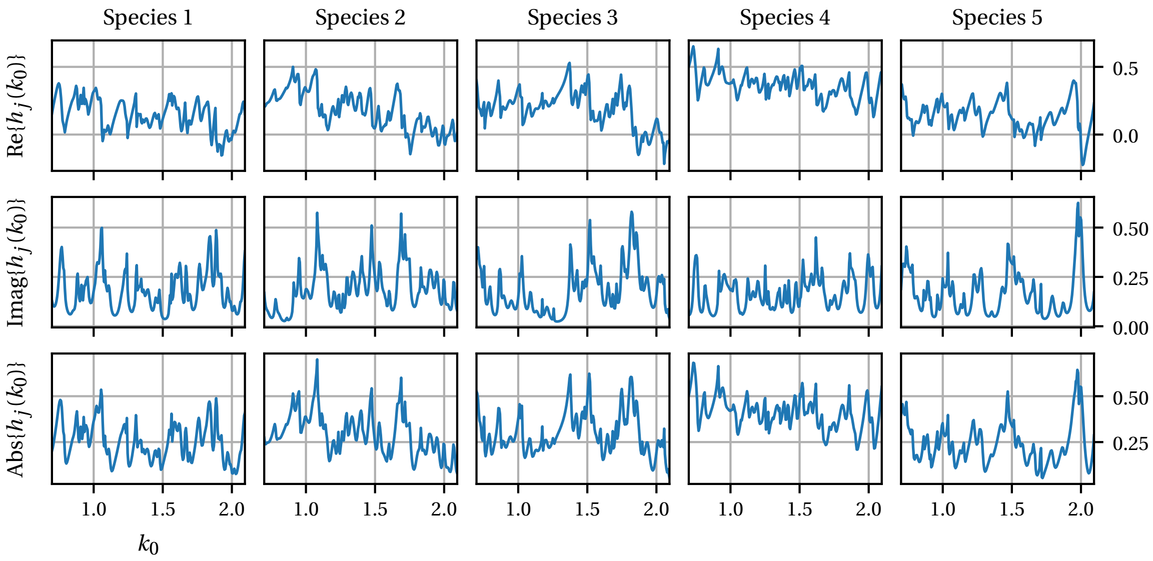


Figure 5.13: Spectral profiles for cell phantom, plotted for $\delta = 1$.

expected, only the edges of the phantom that are nearly perpendicular to the optical axis are visible. The reconstructed images deteriorate as δ increases, particularly at the rear edge of each feature. However, the correct species is identified in each case; negligible energy is deposited into Species 4 and 5. Figure 5.15 maps each species into a single channel of an RGB image. The edges are assigned to the correct species.

Figure 5.16 illustrates the influence of the regularization parameter λ_r. Noise dominates the reconstruction when λ_r is too small. When λ_r is too large, there is no chemical identification-the recovered spatial densities are nearly identical for each species.

5.7 Conclusions

We have considered the problem of chemically specific and spatially resolved tomographic imaging from interferometric measurements. We require the target to be the linear combination of a finite number of distinct chemical species given data at a small number of *en-face* focal planes.

We developed necessary and sufficient conditions for unique recovery of a target satisfying this model. Linear independence of the chemical spectra is not sufficient—additional spectral diversity is required.

In this chapter, we assume the chemical spectra were either known or drawn from a library of possible spectra. We proved that in the latter case, the number of required focal planes scales

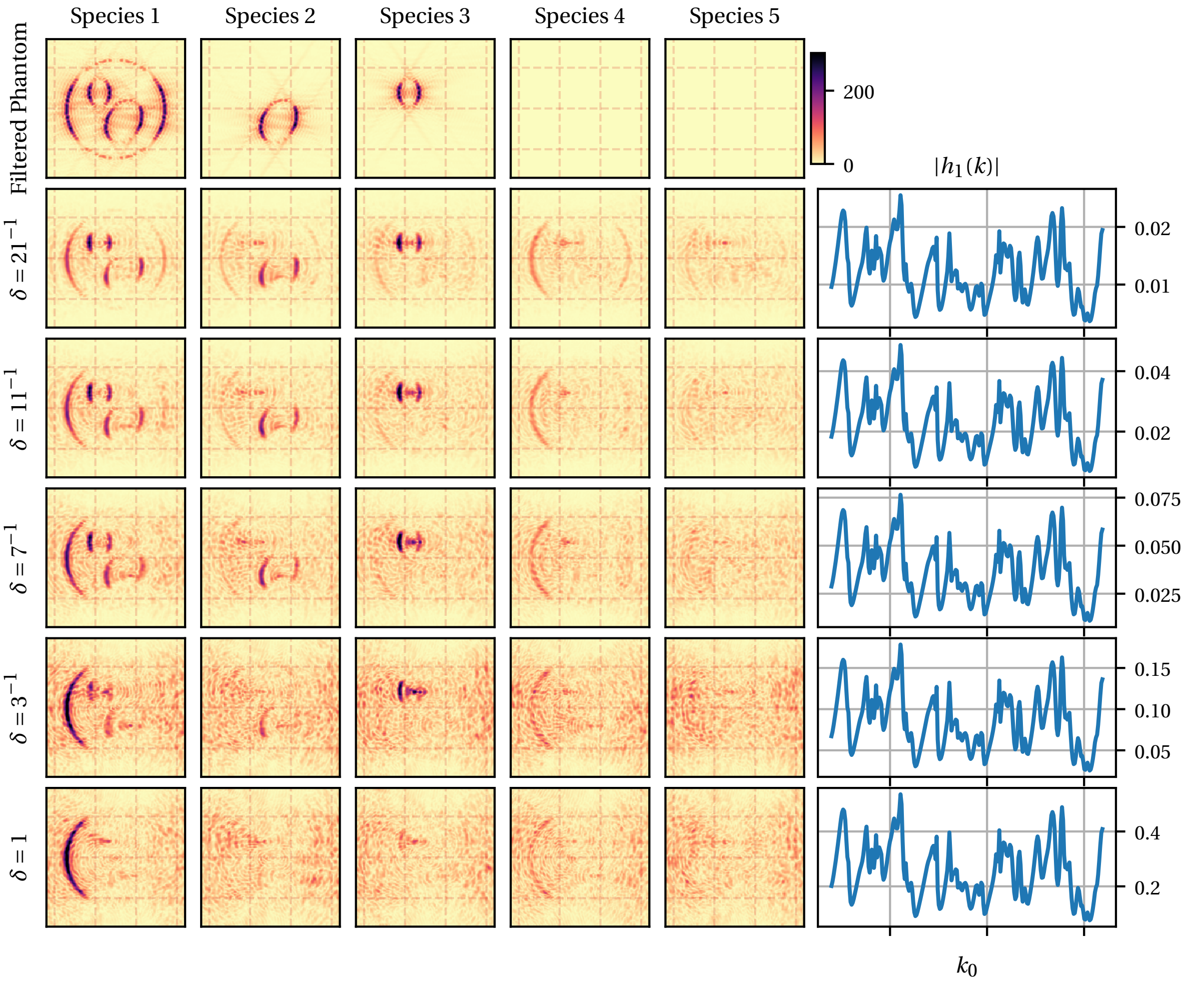


Figure 5.14: Reconstructions of cell phantom as a function of scattering strength. All reconstructions use Tikhonov regularization with $\lambda_r = 10^{-5}$.

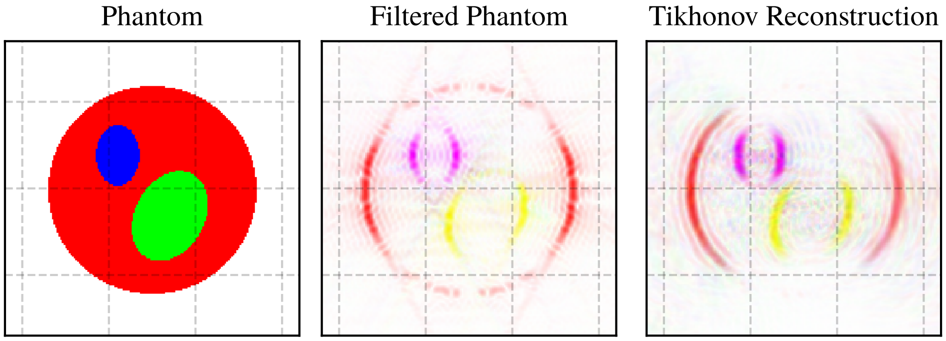


Figure 5.15: Visualizing the reconstructed cell phantom targets by assigning each of the reconstructed species to one RGB channel.

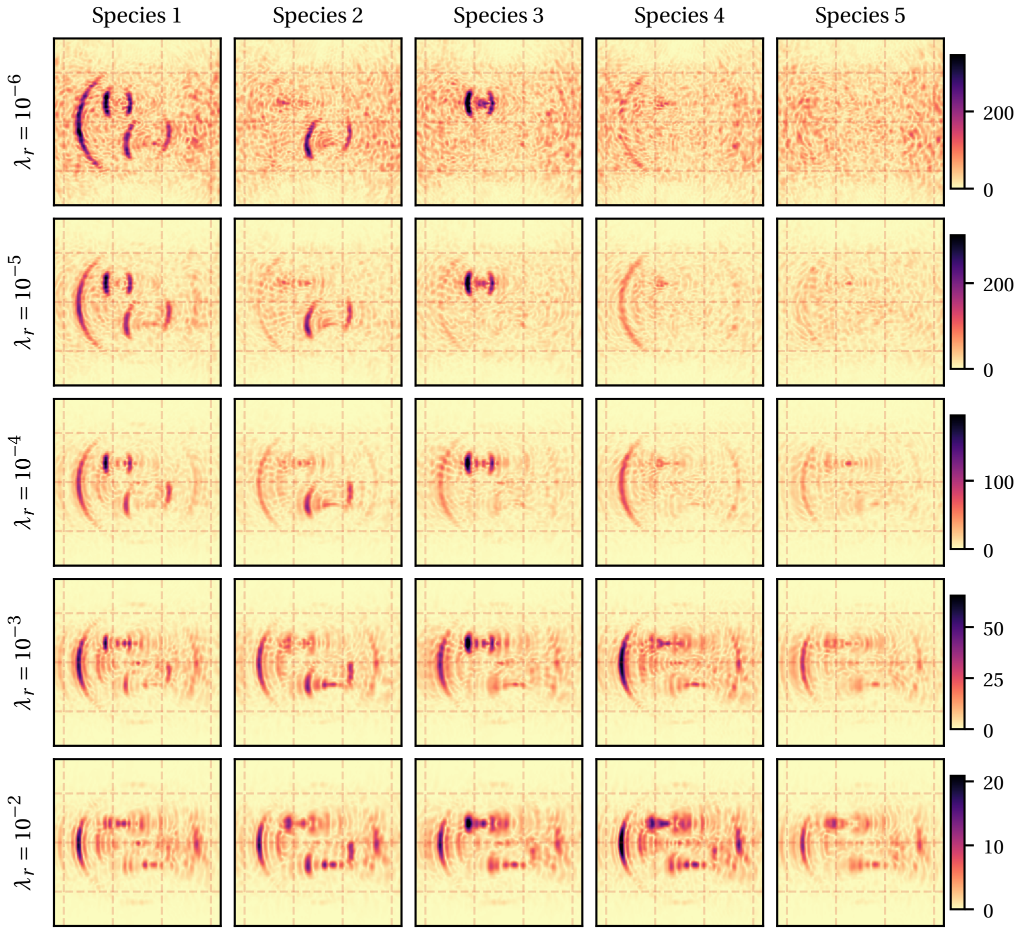


Figure 5.16: Reconstructions of cellular phantom using Tikhonov regularization and various levels of λ_r. The scattering strength parameter is $\delta = 11^{-1}$. As λ_r increases, the reconstruction fails to distinguish between chemical species.

with the number of chemicals present in the sample, not the total number in the library. Future work will consider extension fully blind problem.

Our approach requires interferometric (phase-resolved) measurements and solves the linearized scattering problem. This extension to intensity-only measurements and the removal of the Born approximation are two avenues for future work.

Phaseless, intensity-only diffraction tomography has been demonstrated by modifying the acquisition scheme [168, 192, 193] and by optimization-based approaches [194]. Advances in high performance computing [188, 189, 195] and deep learning [196–198] have facilitated the solution of large scale inverse scattering problems without linearization. In some cases, solving the nonlinear inverse scattering problem overcomes the "missing cone" effect that hampers our reconstruction of extended targets. However, thus far, these approaches have only considered non-dispersive objects. Extension of these methods to spectroscopic tomography within the N-species approximation is an exciting area of future work.

Appendix A

Proofs from Chapter 3

A.1 Proof of Proposition 2.1

We explicitly show the link between filter banks and applying a sparsifying transform to a patch matrix. We assume a 1D signal $x \in \mathbb{R}^N$ to simplify notation. The extension to multiple dimensions is tedious, but straightforward.

Proof. Let $W \in \mathbb{R}^{N_c \times K}$ be a given transform, and let w^i indicate the i-th row of this matrix. Suppose we extract patches with a patch stride of s and we assume s evenly divides N. The j-th column of the patch matrix $X \in \mathbb{R}^{K \times M}$ is the vector $[x_{sj+K-1}, x_{sj+K-2}, \ldots, x_{sj}]^T$. The number of columns, M, depends on the boundary conditions used. Linear and circular convolution are obtained by setting $x_i = 0$ or $x_i = x_{N-i-1}$, respectively, when $i < 0$. For cyclic convolution, we have $M = N/s$. The i, j-th element of the sparsified signal WX is

$$\begin{aligned}
[WX]_{i,j} &= \sum_{k=1}^{K} W_{i,k} X_{k,j} = \sum_{k=1}^{K} W_{i,k} x_{sj+K-1-i} \\
&= (w^i * x)[sj + K - 1].
\end{aligned}$$

Thus the i-th row of WX is the convolution between the filter with impulse response w^i and signal x, followed by downsampling by a factor of s, and shifted by $K-1$. The filter bank has N_c channels with impulse responses given by the rows of W. The shift of $K-1$ can be incorporated into the definition of the patch extraction procedure. For 1D signals, the "first" patch should be $[x_{K-1}, \ldots x_0]^T$, while for 2D signals, the lower-right pixel of the "first" patch is $x[0,0]$. □

A.2 Proof of Proposition 2.2

Proof. The function $J_1(W)$ in (2.8) acts only on the magnitude responses of the filters in $\mathcal{H}$. Let $V \triangleq \left|\bar{F} W^T\right|^2 \in \mathbb{R}^{N_F^2 \times N_c}$. The sum of the i-th column of V is equal to the norm of the i-th filter and, by Lemma 2.1, the eigenvalues of $\mathcal{H}^*\mathcal{H}$ are equal to the row sums of V. Thus, V is generated by a UNTF if and only if the row sums and column sums are constant.

Let $V^\sharp$ be a stationary point of J_1. For each $1 \leq r \leq N_F^2$ and $1 \leq s \leq N_c$, we have

$$\frac{\partial}{\partial V_{r,s}} J_1(V^\sharp) = \frac{1}{2} - \frac{1}{\sum_{j=1}^{N_c} V_{r,j}^\sharp} - \frac{1}{\sum_{i=1}^{N_F^2} V_{i,s}^\sharp} = 0. \tag{A.1}$$

Note that $J_1(V) = +\infty$ if either a row or column of V is identically zero, so $V^\sharp$ is a minimizer only if there is at least one non-zero in each row and column of $V^\sharp$. Subtracting $\frac{\partial}{\partial V_{r',s}} J_1(V^\sharp)$ from $\frac{\partial}{\partial V_{r,s}} J_1(V^\sharp)$ yields $\sum_{j=1}^{N_c} V_{r,j}^\sharp = \sum_{j=1}^{N_c} V_{r',j}^\sharp \triangleq a$. Similarly, subtracting $\frac{\partial}{\partial V_{r,s}} J_1(V^\sharp)$ from $\frac{\partial}{\partial V_{r,s'}} J_1(V^\sharp)$ yields $\sum_{i=1}^{N_F^2} V_{i,s}^\sharp = \sum_{i=1}^{N_F^2} V_{i,s'}^\sharp \triangleq b$. As the row and column sums are uniform for each r and s, we conclude $V^\sharp$ is a UNTF. Next, we have

$$\begin{aligned} \sum_{i=1}^{N_F^2} \sum_{j=1}^{N_c} V_{i,j}^\sharp &= \sum_{i=1}^{N_F^2} \left(\sum_{j=1}^{N_c} V_{i,j}^\sharp \right) = N_F^2 a \\ &= \sum_{j=1}^{N_C} \left(\sum_{i=1}^{N_F^2} V_{i,j}^\sharp \right) = N_c b, \end{aligned}$$

from which we conclude $b = \frac{N_F^2}{N_c} a$. Substituting into (A.1), we find

$$a = 2\left(1 + \frac{N_c}{N_F^2}\right), \qquad b = 2\left(\frac{N_F^2}{N_c} + 1\right),$$

and this completes the proof. □

Appendix B

Proofs from Chapter 5

B.1 Lemmata

Lemma B.1. *Let $d \in L^2(I)$ be a bounded, real valued, positive, and continuous function over the (possibly infinite) interval $I = [a,b]$. Define the "diagonal" multiplication operator $\mathcal{D} : L^2(I) \to L^2(I)$ by $(\mathcal{D}f)(x) = d(x)f(x)$. Then*

$$\min_{x\in I} d(x)\cdot \|f\|_{L^2(I)} \le \|\mathcal{D}f\|_{L^2(I)} \le \max_{x\in I} d(x)\cdot \|f\|_{L^2(I)}.$$

Moreover, $\mathcal{D}^{-1}$ exists and is given by $(\mathcal{D}^{-1}g)(x) = g(x)/d(x)$, and

$$\frac{\|f\|_{L^2(I)}}{\max_{x\in I} d(x)} \le \|\mathcal{D}^{-1}f\|_{L^2(I)} \le \frac{\|f\|_{L^2(I)}}{\min_{x\in I} d(x)}.$$

Theorem B.2 (Variational Characterization of Eigenvalues of a Compact, Self-Adjoint Operator)**.** *Let $\mathcal{A} : \mathsf{H} \to \mathsf{H}$ be a compact, self-adjoint operator on a Hilbert space H equipped with inner product $\langle\cdot,\cdot\rangle$. Let $\lambda_1 \ge \lambda_2 \ge \ldots$ be the eigenvalues of $\mathcal{A}$ listed in non-increasing order and repeated with multiplicity. Assume that at least n eigenvalues exist. Let U be a subspace of H. Then*

$$\lambda_n = \max_{\dim \mathsf{U}=n} \min_{\substack{x\in \mathsf{U}\\ x\neq 0}} \frac{\langle \mathcal{A}x, x\rangle}{\langle x, x\rangle} = \min_{\dim \mathsf{U}=n-1} \max_{\substack{x\in \mathsf{U}^\perp\\ x\neq 0}} \frac{\langle \mathcal{A}x, x\rangle}{\langle x, x\rangle}.$$

Lemma B.3. *The trancendental equation*

$$\tau = \frac{\mu \mathrm{NA}^2}{4}\tan\left(\tau\frac{k_a - k_b}{2}\right) \tag{B.1}$$

has exactly one solution in the interval $\left[\frac{\pi(2n+1)}{k_b-k_a}, \frac{2\pi(n+1)}{k_b-k_a}\right]$, and

$$\tau = -\frac{\mu \mathrm{NA}^2}{4}\cot\left(\tau\frac{k_a - k_b}{2}\right)$$

has exactly one solution in the interval $\left[\frac{2\pi n}{k_b-k_a}, \frac{\pi(2n+1)}{k_b-k_a}\right]$.

Proof. We prove the first claim; the second follows from similar arguments.

Let n be a positive integer. Recall μ, NA^2 are positive. We have

$$\lim_{\tau\to\frac{2\pi(n+1/2)}{k_b-k_a}} \tan\left(\tau\frac{k_a-k_b}{2}\right) = \infty$$
$$\lim_{\tau\to\frac{2\pi(n+1)}{k_b-k_a}} \tan\left(\tau\frac{k_a-k_b}{2}\right) = 0,$$

and $\frac{\mu\mathrm{NA}^2}{4}\tan\left(\tau\frac{k_a-k_b}{2}\right)$ is continuous and monotone decreasing for $\frac{\pi(2n+1)}{k_b-k_a} \le \tau \le \frac{2\pi(n+1)}{k_b-k_a}$. As the function $\tau \mapsto \tau$ is monotone increasing, there is one solution to (B.1) on this interval. □

Lemma B.4. *Let* $-\infty < a < b < \infty$. *Let* $\mathcal{G} : L^2[a,b] \to L^2[a,b]$ *be a self-adjoint and Hilbert-Schmidt. Let* $d \in L^2([a,b])$ *be a real valued, bounded, positive, and continuous function. Define* $\mathcal{D} : L^2[a,b] \to L^2[a,b]$ *by* $(\mathcal{D}f)(x) = d(x)f(x)$.

Let the eigenvalues of $\mathcal{G}$ *be* $\lambda_1 \ge \lambda_2 \ldots$, *listed in non-increasing order and repeated with multiplicity, and let* $\gamma_1 \ge \gamma_2 \ldots$ *be the eigenvalues of* $\mathcal{DGD}$ *be listed in an identical fashion. Then* γ_n *satisfies*

$$\lambda_n \cdot \left(\min_{x\in[a,b]} d(x)\right)^2 \le \gamma_n \le \lambda_n \cdot \left(\max_{x\in[a,b]} d(x)\right)^2.$$

Proof. Let $\langle\cdot,\cdot\rangle$ denote the inner product on $L^2[a,b]$ and set $\|x\|^2 = \langle x, x\rangle$. Using Lemma B.1 and Theorem B.2, we have

$$\begin{aligned}\gamma_n &= \max_{\dim \mathrm{U}=n} \min_{\substack{x\in \mathrm{U}\\ x\neq 0}} \frac{\langle \mathcal{DGD}x, x\rangle}{\|x\|^2} = \max_{\dim \mathrm{U}=n} \min_{\substack{x\in \mathrm{U}\\ x\neq 0}} \frac{\langle \mathcal{GD}x, \mathcal{D}x\rangle}{\|x\|^2}\\ &= \max_{\dim \mathrm{U}=n} \min_{\substack{v\in \mathrm{U}\\ v\neq 0}} \frac{\langle \mathcal{G}v, v\rangle}{\|\mathcal{D}^{-1}v\|^2}\\ &\ge \max_{\dim \mathrm{U}=n} \min_{\substack{v\in \mathrm{U}\\ v\neq 0}} \frac{\langle \mathcal{G}v, v\rangle}{\|v\|^2}\left(\min_{x\in[a,b]} d(x)^2\right)\\ &= \lambda_n\left(\min_{x\in[a,b]} d(x)^2\right).\end{aligned}$$

For the upper bound,

$$\begin{aligned}\gamma_n &= \min_{\dim \mathsf{U}=n-1} \max_{\substack{x\in \mathsf{U}^\perp \\ x\neq 0}} \frac{\langle \mathcal{D}\mathcal{G}\mathcal{D}x, x\rangle}{\|x\|^2} = \min_{\dim \mathsf{U}=n-1} \max_{\substack{x\in \mathsf{U}^\perp \\ x\neq 0}} \frac{\langle \mathcal{G}\mathcal{D}x, \mathcal{D}x\rangle}{\|\mathcal{D}x\|^2} \frac{\|\mathcal{D}x\|^2}{\|x\|^2} \\ &\leq \min_{\dim \mathsf{U}=n-1} \max_{\substack{x\in \mathsf{U}^\perp \\ x\neq 0}} \frac{\langle \mathcal{G}\mathcal{D}x, \mathcal{D}x\rangle}{\|\mathcal{D}x\|^2} \left(\max_{x\in[a,b]} d(x)^2\right) \\ &= \lambda_n \left(\max_{x\in[a,b]} d(x)^2\right).\end{aligned}$$

□

Lemma B.5. *Let $a, b \in \mathbb{R}$. Two linearly independent solutions to*

$$-x^2 f'' - x f' + \left(\frac{a^2}{4} - \frac{b^2}{4}x\right) f = 0$$

are

$$J\left(a, b\sqrt{x}\right),$$

where $J(a, x)$ is the Bessel function of the first kind and of order a, and

$$Y\left(a, b\sqrt{x}\right),$$

where $Y(a, x)$ is the Bessel function of the second kind and of order a.

Proof. The change of variables $u = b\sqrt{x}$ yields Bessel's equation,

$$-u^2 \frac{\mathrm{d}^2 f}{\mathrm{d}u^2} - u\frac{\mathrm{d}f}{\mathrm{d}u} + \left(a^2 - u^2\right) f = 0,$$

which has linearly independent solutions $J(a, u)$ and $Y(a, u)$. □

B.2 Proof of Proposition 4.1

Proof. First, set $\sigma = \|w\|_{L^2[k_a, k_b]}$; we have

$$\sigma^2 \triangleq \int_{k_a}^{k_b} |w(k_0)|^2 \,\mathrm{d}k_0 = \frac{4\mu\pi}{\gamma}\left(\frac{1}{k_a+\mu} - \frac{1}{k_b+\mu}\right) = \frac{4\pi\mu}{\gamma}\frac{k_b - k_a}{(k_a+\mu)(k_b+\mu)}.$$

Expanding $\|W - wv\|^2_{L^2([k_a,k_b]\times\mathbb{R})}$ and using (4.5) and $\|v\|_{L^2(\mathbb{R})} = 1$,

$$\begin{aligned}
\|W - wv\|^2_{L^2[k_a,k_b]\times\mathbb{R})} &= \int_{-\infty}^{\infty}\int_{k_a}^{k_b} |W(k_0,z) - w(k_0)v(z)|^2 \,\mathrm{d}k_0\mathrm{d}z \\
&= \|W\|^2_{L^2([k_a,k_b]\times\mathbb{R})} + \|w\|^2_{L^2([k_a,k_b])}\|v\|^2_{L^2(\mathbb{R})} - 2\mathrm{Re}\left(\int_{k_a}^{k_b}\int_{-\infty}^{\infty} w(k_0)W^*(k_0,z)v(z)\mathrm{d}z\mathrm{d}k_0\right) \\
&= \frac{\pi}{\gamma}\log\left(\frac{k_b}{k_a}\right) + \sigma^2 - 2\mathrm{Re}\left(\int_{k_a}^{k_b} |w(k_0)|^2\,\mathrm{d}k_0\right) \\
&= \frac{\pi}{\gamma}\log\left(\frac{k_b}{k_a}\right) - \sigma^2,
\end{aligned}$$

thus we can choose μ to maximize σ instead. To that end,

$$\frac{\partial}{\partial\mu}\sigma^2 = \frac{4\pi}{\gamma}\frac{(k_b-k_a)(k_ak_b-\mu^2)}{(k_a+\mu)^2(k_b+\mu)^2},$$

thus $\mu^\star = \sqrt{k_ak_b}$ is a stationary point. Additionally,

$$\begin{aligned}
\frac{\partial^2}{\partial\mu^2}\sigma^2\Big|_{\mu=\mu^\star} &= \frac{8\pi}{\gamma}\frac{(k_a-k_b)(k_a^2k_b-(\mu^\star)^3+k_ak_b(k_b+3\mu^\star))}{(k_a+\mu^\star)^3(k_b+\mu^\star)^3} \\
&= \frac{8\pi}{\gamma}\frac{(k_a-k_b)(k_a+k_b+2\sqrt{k_ak_b})}{(k_a+\sqrt{k_ak_b})^3(k_b+\sqrt{k_ak_b})^3} < 0,
\end{aligned}$$

thus $\mu^\star$ maximizes σ^2. □

B.3 Proof of Proposition 4.2

Proof. The kernel of $\mathcal{G}_{\mathbf{k}_\parallel}$ is real valued and symmetric; thus $\mathcal{G}_{\mathbf{k}_\parallel}$ is self-adjoint. Moreover, we have

$$\begin{aligned}
\int_{k_a}^{k_b}\int_{k_a}^{k_b}\left|G_{\mathbf{k}_\parallel}(k,k')\right|^2\mathrm{d}k\mathrm{d}k' &= \int_{k_a}^{k_b}\int_{k_a}^{k_b}\left|\exp\left\{\frac{-\left|k-k'\right|}{\gamma\mu}\left(2+\frac{\left|\mathbf{k}_\parallel\right|^2}{4kk'}\right)\right\}\right|^2\mathrm{d}k\mathrm{d}k' \\
&\le \int_{k_a}^{k_b}\int_{k_a}^{k_b}\exp\left\{\frac{-4\left|k-k'\right|}{\gamma\mu}\right\}\mathrm{d}k\mathrm{d}k' \\
&= \left(\frac{\gamma\mu}{2}(k_b-k_a) + \frac{\gamma^2\mu^2}{8}\left(e^{\frac{-4}{\gamma\mu}(k_b-k_a)}-1\right)\right) \\
&< \infty,
\end{aligned}$$

and thus $\mathcal{G}_{\mathbf{k}_\parallel}$ is Hilbert-Schmidt. As $\mathcal{D}_{\mathbf{k}_\parallel}$ and $\mathcal{G}_{\mathbf{k}_\parallel}$ are self-adjoint, $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}\tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel} = \mathcal{D}_{\mathbf{k}_\parallel}\mathcal{G}_{\mathbf{k}_\parallel}\mathcal{D}_{\mathbf{k}_\parallel}$ is self-adjoint. Further, the kernel of $\tilde{\mathcal{A}}_{\mathbf{k}_\parallel}\tilde{\mathcal{A}}^*_{\mathbf{k}_\parallel}$ satisfies

$$\begin{aligned}\int_{k_a}^{k_b}\int_{k_a}^{k_b}\left|(\tilde{A}_{\mathbf{k}_\parallel}\tilde{A}_{\mathbf{k}_\parallel})(k,k')\right|^2 \mathrm{d}k\mathrm{d}k' &= \int_{k_a}^{k_b}\int_{k_a}^{k_b}\left| d_{\mathbf{k}_\parallel}(k)\exp\left\{\frac{-|k-k'|}{\gamma\mu}\left(2+\frac{|\mathbf{k}_\parallel|^2}{4kk'}\right)\right\} d_{\mathbf{k}_\parallel}(k')\right|^2 \mathrm{d}k\mathrm{d}k' \\ &\leq \|d_{\mathbf{k}_\parallel}\|_\infty^4 \iint_{k_a}^{k_b}\left|G_{\mathbf{k}_\parallel}(k,k')\right|^2 \mathrm{d}k\mathrm{d}k' \\ &< \infty,\end{aligned}$$

and thus the operator is Hilbert-Schmidt. □

B.3.1 Proof of Theorem 4.1

Proof. Let f be an eigenfunction of $\mathcal{G}_{\mathbf{k}_\parallel}$ with eigenvalue $\lambda > 0$. We will use direct computation to show that the pair (f,λ) satisfy (4.14) to (4.16).

We begin by introducing some additional notation. Define the scalars

$$\begin{aligned}\delta &\triangleq \frac{2}{\gamma\mu} = \frac{4}{\mathrm{NA}^2\mu} \\ \tau &\triangleq \frac{|\mathbf{k}_\parallel|^2}{4\gamma\mu} = \frac{|\mathbf{k}_\parallel|^2}{2\mu\mathrm{NA}^2},\end{aligned}$$

so that $\tilde{G}_{\mathbf{k}_\parallel}(k,t)$ in (4.10) and $p(k)$ in (4.13) can be written

$$\begin{aligned}\tilde{G}_{\mathbf{k}_\parallel}(k,t) &= e^{-\delta|k-\mathbf{k}_\parallel|+\tau|k^{-1}-t^{-1}|} \\ p(k) &= \frac{k^2}{\delta k^2+\tau}.\end{aligned}$$

It is convenient to work with $q = 1/p$; explicitly,

$$q(k) \triangleq \frac{1}{p(k)} = \delta + \frac{\tau}{k^2}.$$

Define the functions

$$\begin{aligned}L_1(k,t) &\triangleq e^{\delta(t-k)+\tau(k^{-1}-t^{-1})} \\ L_2(k,t) &\triangleq e^{\delta(k-t)+\tau(t^{-1}-k^{-1})}\end{aligned}$$

so that the kernel of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ (4.10) can be written in piecewise form as

$$\begin{aligned}\tilde{G}_{\mathbf{k}_\parallel}(k,t) &= e^{-\delta|k-t|-\tau\left|\frac{1}{k}-\frac{1}{t}\right|}\\ &= L_1(k,t)\cdot\mathbb{1}_{t\le k} + L_2(k,t)\cdot\mathbb{1}_{t>k}.\end{aligned}$$

Note that $L_1(k,k) = L_2(k,k)$ and $\tilde{G}_{\mathbf{k}_\parallel}(k,t)$ is continuous in k and t. The following derivatives will be useful:

$$\begin{aligned}\frac{\partial}{\partial k}L_1(k,t) &= -q(k)L_1(k,t)\\ \frac{\partial}{\partial k}L_2(k,t) &= q(k)L_2(k,t).\end{aligned}$$

Let $g = \tilde{\mathcal{G}}_{\mathbf{k}_\parallel} f$; explicitly,

$$g(k) = \int_{k_a}^{k_b} \tilde{G}_{\mathbf{k}_\parallel}(k,t)f(t)\mathrm{d}t = \int_{k_a}^{k} L_1(k,t)f(t)\mathrm{d}t + \int_{k}^{k_b} L_2(k,t)f(t)\mathrm{d}t. \tag{B.2}$$

Next, we must calculate g' and g''. Differentiating with respect to k, we have

$$\begin{aligned}g'(k) &= \frac{\mathrm{d}}{\mathrm{d}k}\int_{k_a}^{k} L_1(k,t)f(t)\mathrm{d}t + \frac{\mathrm{d}}{\mathrm{d}k}\int_{k}^{k_b} L_2(k,t)f(t)\mathrm{d}t\\ &= L_1(k,k)f(k) + \int_{k_a}^{k}\frac{\partial}{\partial k}L_1(k,t)f(t)\mathrm{d}t - L_2(k,k)f(k) + \int_{k}^{k_b}\frac{\partial}{\partial k}L_2(k,t)f(t)\mathrm{d}t\\ &= -q(k)\int_{k_a}^{k} L_1(k,t)f(t)\mathrm{d}t + q(k)\int_{k}^{k_b} L_2(k,t)f(t)\mathrm{d}t. \qquad \text{(B.3)}\end{aligned}$$

We compute the derivative of each term in (B.3) independently;

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}k}\left(q(k)\int_{k_a}^{k} L_1(k,t)f(t)\mathrm{d}t\right) &= -\frac{2\tau}{k^3}\int_{k_a}^{k} L_1(k,t)f(t)\mathrm{d}t + q(k)\frac{\mathrm{d}}{\mathrm{d}k}\int_{k_a}^{k} L_1(k,t)f(t)\mathrm{d}t\\ &= -\frac{2\tau}{k^3}\int_{k_a}^{k} L_1(k,t)f(t)\mathrm{d}t + q(k)\left(f(k) - q(k)\int_{k_a}^{k} L_1(k,t)f(t)\mathrm{d}t\right)\\ &= q(k)f(k) - \left(\frac{2\tau}{k^3} + q(k)^2\right)\int_{k_a}^{k} L_1(k,t)f(t)\mathrm{d}t \qquad \text{(B.4)}\\ \frac{\mathrm{d}}{\mathrm{d}k}\left(q(k)\int_{k}^{k_b} L_2(k,t)f(t)\mathrm{d}t\right) &= -\frac{2\tau}{k^3}\int_{k}^{k_b} L_2(k,t)f(t)\mathrm{d}t + q(k)\frac{\mathrm{d}}{\mathrm{d}k}\int_{k}^{k_b} L_2(k,t)f(t)\mathrm{d}t\\ &= -\frac{2\tau}{k^3}\int_{k}^{k_b} L_2(k,t)f(t)\mathrm{d}t + q(k)\left(-f(k) + q(k)\int_{k}^{k_b} L_2(k,t)f(t)\mathrm{d}t\right)\\ &= -q(k)f(k) + \left(-\frac{2\tau}{k^3} + q(k)^2\right)\int_{k_a}^{k} L_2(k,t)f(t)\mathrm{d}t. \qquad \text{(B.5)}\end{aligned}$$

Combining (B.4) and (B.5) and simplifying using (B.3) yields

$$\begin{aligned} g''(k) &= \frac{\mathrm{d}}{\mathrm{d}k}\left(q(k)\int_{k_a}^{k} L_1(k,t)f(t)\mathrm{d}t\right) + \frac{\mathrm{d}}{\mathrm{d}k}\left(q(k)\int_{k}^{k_b} L_2(k,t)f(t)\mathrm{d}t\right) \\ &= -2q(k)f(k) + q(k)^2 g(k) - \frac{2\tau}{k^3}\frac{1}{q(k)}g'(k). \end{aligned} \tag{B.6}$$

By assumption f is an eigenfunction of $\tilde{\mathcal{G}}_{\mathbf{k}_\parallel}$ with eigenvalue λ. Thus $g = \tilde{\mathcal{G}}_{\mathbf{k}_\parallel} f = \lambda f, g' = \lambda f'$, and $g'' = \lambda f''$. Thus by (B.6), f'' must satisfy the differential equation

$$\lambda f''(k) = -2q(k)f(k) + \lambda q(k)^2 f(k) - \frac{2\tau}{k^3}\frac{1}{q(k)}\lambda f'(k).$$

Recall that $p(k) = 1/q(k)$. Simplifying and rearranging yields,

$$-p(k)f''(k) - \frac{2\tau}{k^3}p(k)^2 f'(k) + \frac{f(k)}{p(k)} = \frac{2}{\lambda}f(k). \tag{B.7}$$

Note $p'(k) = 2\tau p^2(k)/k^3$; thus (B.7) is equivalent to

$$-p(k)f''(k) - p'(k)f'(k) + \frac{f(k)}{p(k)} = \frac{2}{\lambda}f(k),$$

which is easily converted to the desired Sturm-Liouville equation

$$-\frac{\mathrm{d}}{\mathrm{d}k}\big(p(k)f'(k)\big) - \frac{f(k)}{p(k)} = \frac{2}{\lambda}f(k).$$

Finally, we show that f satisfies the mixed boundary conditions (4.15) and (4.16). Recall $g = \lambda f$, $g' = \lambda f'$, and $p = 1/q$. By (B.2) and (B.3) we have

$$\lambda f(k) = \int_{k_a}^{k} L_1(k,t)f(t)\mathrm{d}t + \int_{k}^{k_b} L_2(k,t)f(t)\mathrm{d}t \tag{B.8}$$

$$\lambda p(k)f'(k) = -\int_{k_a}^{k} L_1(k,t)f(t)\mathrm{d}t + \int_{k}^{k_b} L_2(k,t)f(t)\mathrm{d}t. \tag{B.9}$$

Evaluating (B.8) and (B.9) at $k = k_a$ yields

$$\begin{aligned} \lambda f(k_a) &= \int_{k_a}^{k_b} L_2(k_a,t)f(t)\mathrm{d}t \\ \lambda f'(k_a) &= \frac{1}{p(k_a)}\int_{k_a}^{k_b} L_2(k_a,t)f(t)\mathrm{d}t. \end{aligned}$$

thus

$$\lambda\left(f(k_a) - p(k_a)f'(k_a)\right) = 0,$$

and as $\lambda \neq 0$, we must have $f(k_a) - p(k_a)f'(k_a) = 0$, establishing (4.15). Similarly, evaluating (B.8) and (B.9) at $k = k_b$ yields

$$\begin{aligned}\lambda f(k_b) &= \int_{k_a}^{k_b} L_2(k_b,t)f(t)\mathrm{d}t\\ \lambda f'(k_b) &= -\frac{1}{p(k_b)}\int_{k_a}^{k_b} L_2(k_b,t)f(t)\mathrm{d}t\end{aligned}$$

and thus

$$\lambda\left(f(k_b) - p(k_b)f'(k_b)\right) = 0,$$

establishing the boundary condition (4.16). □

B.4 Proof of Theorem 4.3

Proof. Suppose f is an eigenfunction of $\tilde{\mathcal{G}}_0$ with non-zero eigenfunction λ. Set

$$\delta = \frac{1}{p(k)} = \frac{4}{\mathrm{NA}^2\mu}$$

and rewrite the ODE (4.18) as

$$-f'' + \left(\delta^2 - \frac{2\delta}{\lambda}\right)f = 0 \tag{B.10}$$

subject to boundary conditions

$$f(k_a) - \delta^{-1}f'(k_a) = 0 \tag{B.11}$$

$$f(k_b) + \delta^{-1}f'(k_b) = 0. \tag{B.12}$$

First, we show that $\delta^2 - 2\delta/\lambda < 0$. Let λ_1 be the largest eigenvalue of $\tilde{\mathcal{G}}_0$. Specialized to the problem at hand, the inequality (4.17) can be written

$$1 \leq \frac{\delta}{4}(k_b - k_a)\left(\int_{k_a}^{k_b} \max\left\{\frac{2}{\lambda_1} - \delta, 0\right\}\mathrm{d}k\right),$$

which is non-zero only if $\lambda_1 \leq 2/\delta$. As $0 < \lambda < \lambda_1$, we have $\delta^2 - 2\delta^2/\lambda < 0$, as desired.

Define $k_{\mathrm{mid}} = (k_a + k_b)/2$ and set

$$\tau \triangleq \sqrt{\frac{2\delta}{\lambda} - \delta^2}. \tag{B.13}$$

Two linearly independent solutions to (B.10) $\cos(\tau(k - k_{\mathrm{mid}}))$ and $\sin(\tau(k - k_{\mathrm{mid}}))$, so any f satisfying (B.10) can be written in the form $f(k) = A\cos(\tau(k - k_{\mathrm{mid}})) + \sin(\tau(k - k_{\mathrm{mid}}))$. Next, we show that f satisfying the boundary conditions (B.11) and (B.12) and $A \neq 0$ must have $B = 0$ and vice versa.

Suppose $A \neq 0$; without loss of generality we can take $A = 1$. The boundary conditions (B.11) and (B.12) become

$$\begin{aligned}(\delta - B\tau)\cos\left(\frac{\tau}{2}(k_a - k_b)\right) + (B\delta + \tau)\sin\left(\frac{\tau}{2}(k_a - k_b)\right) &= 0 \\ (\delta + B\tau)\cos\left(\frac{\tau}{2}(k_a - k_b)\right) + (-B\delta + \tau)\sin\left(\frac{\tau}{2}(k_a - k_b)\right) &= 0.\end{aligned} \tag{B.14}$$

Solving (B.14) for B yields

$$B = \frac{\delta\cos\left(\frac{\tau}{2}(k_a - k_b)\right) + \tau\sin\left(\frac{\tau}{2}(k_a - k_b)\right)}{\tau\cos\left(\frac{\tau}{2}(k_a - k_b)\right) - \delta\sin\left(\frac{\tau}{2}(k_a - k_b)\right)},$$

while solving (B.14) for B yields

$$B = -\frac{\delta\cos\left(\frac{\tau}{2}(k_a - \mu)\right) + \tau\sin\left(\frac{\tau}{2}(k_a - \mu)\right)}{\tau\cos\left(\frac{\tau}{2}(k_a - \mu)\right) - \delta\sin\left(\frac{\tau}{2}(k_a - \mu)\right)}.$$

Thus $B = -B$ and we must have $B = 0$. The same argument for $B = 1$ implies $A = 0$.

Next, we show that τ and δ (and thus λ) are roots of a trancendental equation. Suppose f satisfies (B.10) to (B.12) with $A = 1$. As f satisfies the boundary conditions, we have

$$\delta\cos\left(\frac{\tau}{2}(k_a - k_b)\right) + \tau\sin\left(\frac{\tau}{2}(k_a - k_b)\right) = 0,$$

and thus

$$\tau = -\delta\cot\left(\frac{\tau}{2}(k_a - k_b)\right).$$

If $B = 1$, we have

$$\tau\cos\left(\frac{\tau}{2}(k_a - k_b)\right) - \delta\sin\left(\frac{\tau}{2}(k_a - k_b)\right) = 0,$$

and thus

$$\tau = \delta \tan\left(\frac{\tau}{2}(k_a - k_b)\right).$$

Finally, given τ and δ, we obtain λ from (B.13) as

$$\lambda = \frac{2\delta}{\tau^2 + \delta^2} = \frac{8\mu \mathrm{NA}^2}{16 + \tau^2\mu^2\mathrm{NA}^4}.$$

□

B.5 Proof of Theorem 4.5

Proof of Theorem 4.5. Expand (4.20) and simplify:

$$\begin{aligned}
-\tilde{p}(k)\tilde{f}'' - \tilde{p}'(k)\tilde{f}' + \left(\frac{1}{\tilde{p}(k)} - \frac{2}{\tilde{\lambda}}\right)\tilde{f} = 0 \\
-\beta(\alpha + 2\left|\mathbf{k}_\parallel\right|^2 k)\tilde{f}'' - 2\beta\left|\mathbf{k}_\parallel\right|^2 \tilde{f}' + \left(\frac{1}{\beta(\alpha + 2\left|\mathbf{k}_\parallel\right|^2 k)} - \frac{2}{\tilde{\lambda}}\right)\tilde{f} = 0 \\
-(\alpha + 2\left|\mathbf{k}_\parallel\right|^2 k)\tilde{f}'' - 2\left|\mathbf{k}_\parallel\right|^2 \tilde{f}' + \left(\frac{1}{\beta^2(\alpha + 2\left|\mathbf{k}_\parallel\right|^2 k)} - \frac{2}{\beta\tilde{\lambda}}\right)\tilde{f} = 0.
\end{aligned} \tag{B.15}$$

Make the change of variables $u = \alpha + 2\left|\mathbf{k}_\parallel\right|^2 k$. We have

$$\begin{aligned}
\frac{\mathrm{d}\tilde{f}}{\mathrm{d}k} = 2\left|\mathbf{k}_\parallel\right|^2 \frac{\mathrm{d}\tilde{f}}{\mathrm{d}u} \\
\frac{\mathrm{d}^2\tilde{f}}{\mathrm{d}k^2} = 4\left|\mathbf{k}_\parallel\right|^4 \frac{\mathrm{d}^2\tilde{f}}{\mathrm{d}u^2};
\end{aligned}$$

thus (B.15) becomes

$$-u^2\frac{\mathrm{d}^2\tilde{f}}{\mathrm{d}u^2} - u\frac{\mathrm{d}\tilde{f}}{\mathrm{d}u} + \left(\left(\frac{1}{2\beta\left|\mathbf{k}_\parallel\right|^2}\right)^2 - \frac{2}{4\beta\left|\mathbf{k}_\parallel\right|^4\tilde{\lambda}}\right)\tilde{f} = 0. \tag{B.16}$$

The solution to (B.16) is obtained using Lemma B.5 with $a = (\beta\left|\mathbf{k}_\parallel\right|^2)^{-1}$ and $b = \left|\mathbf{k}_\parallel\right|^{-2}\sqrt{\frac{2}{\beta\tilde{\lambda}}}$. □

Appendix C

Proofs from Chapter 6

C.1 Rank Bounds for the Khatri-Rao Product

The following lemma regarding the rank of the Khatri-Rao product will prove useful:

Lemma C.1. *Given* $\mathbf{A} \in \mathbb{C}^{m\times n_1}$ *and* $\mathbf{B} \in \mathbb{C}^{m\times n_2}$*,* $\operatorname{rank}\{\mathbf{A}\odot\mathbf{B}\} \le \min(m, \operatorname{rank}\{\mathbf{A}\}\operatorname{rank}\{\mathbf{B}\})$.

Proof. As $\mathbf{A}\odot\mathbf{B} \in \mathbb{C}^{m\times n_1 n_2}$, we have $\operatorname{rank}\{\mathbf{A}\odot\mathbf{B}\} \le \min(m, n_1 n_2)$. Note that $\mathbf{A}\odot\mathbf{B}$ contains a subset of rows of the matrix $\mathbf{A}\otimes\mathbf{B}$. As the rank of the Kronkecker product is equal to the product of the ranks of $\mathbf{A}$ and $\mathbf{B}$ (*e.g.*, [199]), we have $\operatorname{rank}\{\mathbf{A}\odot\mathbf{B}\} \le \operatorname{rank}\{\mathbf{A}\otimes\mathbf{B}\} = \operatorname{rank}\{\mathbf{A}\}\operatorname{rank}\{\mathbf{B}\}$. □

C.2 Proof of Lemma 5.1

Proof. (C1) $\implies$ (C2): Let $\bar{\mathbf{p}}^l \in \left(\bar{\mathsf{N}}^l\right)^\perp$ be the unique solution to $\bar{\mathbf{s}}^l = \Phi^l\bar{\mathbf{p}}^l$. Let $\mathbf{x} \in \operatorname{null}\left\{\Phi^l\right\} \cap \left(\bar{\mathsf{N}}^l\right)^\perp$. Now $\Phi^l(\bar{\mathbf{p}}^l + \mathbf{x}) = \Phi^l\bar{\mathbf{p}}^l = \bar{\mathbf{s}}^l$. As $\mathbf{x} + \bar{\mathbf{p}}^l \in \left(\bar{\mathsf{N}}^l\right)^\perp$, by (C1) $\mathbf{x} = 0$. Thus (C1) $\implies$ (C2).

(C2) $\implies$ (C3): Recall $\tilde{\Phi}^l = \Phi^l(\mathbf{I}_{N_F} \otimes \mathbf{V}^l) \in \mathbb{C}^{N_F N_k \times N_s r}$. As $\mathbf{I}_{N_F} \otimes \mathbf{V}^l$ is a basis for $\left(\bar{\mathsf{N}}^l\right)^\perp$, and $\operatorname{null}\left\{\Phi^l\right\} = \bar{\mathsf{N}}^l$ by assumption, $\tilde{\Phi}^l\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{x} = \mathbf{0}$; thus $\operatorname{null}\left\{\tilde{\Phi}^l\right\} = \{\mathbf{0}\}$. By the rank nullity theorem, $\operatorname{rank}\left\{\tilde{\Phi}^l\right\} = N_s r$.

(C3) $\implies$ (C1): Suppose $\exists \mathbf{u}, \mathbf{v} \in \left(\bar{\mathsf{N}}^l\right)^\perp$ such that $\Phi^l\mathbf{u} = \Phi^l\mathbf{v}$. As $\mathbf{I}_{N_F} \otimes \mathbf{V}^l$ is a basis for $\left(\bar{\mathsf{N}}^l\right)^\perp$, there are unique vectors $\mathbf{x}, \mathbf{y}$ such that $\mathbf{u} = (\mathbf{I}_{N_F} \otimes \mathbf{V}^l)\mathbf{x}$ and $\mathbf{v} = (\mathbf{I}_{N_F} \otimes \mathbf{V}^l)\mathbf{y}$. Now $\mathbf{0} = \Phi^l(\mathbf{u} - \mathbf{v}) = \tilde{\Phi}^l(\mathbf{x} - \mathbf{y}) \implies \mathbf{x} = \mathbf{y}$ as $\tilde{\Phi}^l$ is full column rank; thus $\mathbf{u} = \mathbf{v}$, completing the proof. □

C.3 Proof of Theorem 5.2

Proof. Here, we suppress the superscript l. By Lemma 5.1, it suffices to show that the proposed conditions are necessary for $\tilde{\Phi}$ to have rank $N_s r$. (N1) follows as $\tilde{\Phi}$ can have rank $N_s r$ only if $N_k N_F \ge N_s r$.

We show (N2) by contradiction; suppose $\operatorname{rank}\{\mathbf{H}\} = q < N_s$. By construction $\operatorname{rank}\{\bar{\mathbf{H}}\} = \operatorname{rank}\{\mathbf{H}\}$. Thus by Lemma C.1, $\operatorname{rank}\{\tilde{\Phi}\} \le \operatorname{rank}\{\bar{\mathbf{H}}\} \operatorname{rank}\{\bar{\mathbf{A}}^l\} \le rq < N_s r$.

For (N3), suppose the first row of $\hat{\mathbf{B}}$ is orthogonal to the remaining $N_k N_F$ rows. Let $\mathbf{x}$ be a column vector formed from first row of $\hat{\mathbf{B}}$ and let $\mathbf{e}_1 \triangleq [1,0,\ldots,0] \in \mathbb{C}^{N_k N_F}$; by construction, $\hat{\mathbf{B}}\mathbf{x} = \mathbf{e}_1$. Set $\alpha = \sum_{j=2}^{N_s} \mathbf{h}_j[1]/\mathbf{h}_1[1]$; then

$$\tilde{\Phi}\, [-\alpha\mathbf{x}^\mathsf{T}, \mathbf{x}^\mathsf{T}, \ldots, \mathbf{x}^\mathsf{T}]^\mathsf{T} = \operatorname{diag}\left\{ \sum_{j=2}^{N_s} \mathbf{h}_j - \alpha\mathbf{h}_1 \right\} \mathbf{e}_1 = \mathbf{0},$$

and so $\operatorname{rank}\{\Phi\} \le N_s r - 1$.

To show (N4), suppose there is a subset J with $|J| \ge N_s$ such that $\mathbf{H}[J,:\,] \in \mathbb{C}^{|J| \times N_s}$ is rank N_s and the remaining rows, $\mathbf{H}[J^c,:\,] \in \mathbb{C}^{N_k - |J| \times N_s}$ has rank $q < N_s$. Define $\tilde{\Phi}^J \in \mathbb{C}^{N_F|J| \times N_s r}$ to be the rows of $\tilde{\Phi}$ involving the rows of $\mathbf{H}$ indexed by J; that is,

$$\tilde{\Phi}^J = \begin{bmatrix} \mathbf{H}[J,:\,] \odot \hat{\mathbf{B}}_1[J,:\,] \\ \vdots \\ \mathbf{H}[J,:\,] \odot \hat{\mathbf{B}}_{N_F}[J,:\,] \end{bmatrix},$$

and construct $\tilde{\Phi}^{J^c} \in \mathbb{C}^{N_F(N_k - |J|) \times N_s r}$ using the rows indexed by J^c. As both $\hat{\mathbf{B}}[J,:\,] \in \mathbb{C}^{N_F|J| \times r}$ and $\hat{\mathbf{B}}[J^c,:\,] \in \mathbb{C}^{N_F(N_k - |J|) \times r}$ have rank at most r, by Lemma C.1, we have

$$\operatorname{rank}\{\tilde{\Phi}\} \le \operatorname{rank}\{\tilde{\Phi}^J\} + \operatorname{rank}\left\{\tilde{\Phi}^{J^c}\right\} \le \min(N_F|J|, N_s r) + \min\left(N_F(N_k - |J|), qr\right) \triangleq \beta.$$

Our goal is establish conditions such that $\beta \ge N_s r$. This is clearly true, regardless of q, when $N_F|J| \ge N_s r$. When $|J| < N_s r / N_F$, we have

$$\beta = N_F|J| + \min\left(N_F(N_k - |J|), qr\right).$$

Suppose $N_F(N_k - |J|) < qr$; then $\beta = N_F N_k \ge N_s r$ where the inequality follows from condition (N1). Otherwise, if $N_F(N_k - |J|) \ge qr$, then $\beta = N_F|J| + qr$ and $q \ge N_s - N_F|J|/r$ implies $\beta \ge N_s r$.

To show (N5), for each $i \in [N_k]$ we define the index set $J_i = \{i, i + N_k, \ldots, i + (N_F - 1)N_k\}$; now,

$$\tilde{\Phi}^{J_i} = (\mathbb{1}_{N_F}^\mathsf{T} \otimes \bar{\mathbf{H}}[J_i,:\,]) \odot \bar{\mathbf{B}}[J_i,:\,] = \begin{bmatrix} \mathbf{h}_1[i]\hat{\mathbf{B}}_1[i,:\,] & \mathbf{h}_2[i]\hat{\mathbf{B}}_1[i,:\,] & \ldots & \mathbf{h}_{N_s}[i]\hat{\mathbf{B}}_1[i,:\,] \\ \vdots & \vdots & & \vdots \\ \mathbf{h}_1[i]\hat{\mathbf{B}}_{N_F}[i,:\,] & \mathbf{h}_2[i]\hat{\mathbf{B}}_{N_F}[i,:\,] & \ldots & \mathbf{h}_{N_s}[i]\hat{\mathbf{B}}_{N_F}[i,:\,] \end{bmatrix} \in \mathbb{C}^{N_F \times N_s r}.$$

Now, $\operatorname{rank}\{\tilde{\Phi}\} \leq \sum_{i=1}^{N_k} \operatorname{rank}\{\tilde{\Phi}^{J_i}\} \leq \sum_{i=1}^{N_k} \operatorname{rank}\{\hat{\mathbf{B}}[J_i, :]\}$, where the final inequality follows from Lemma C.1 and $\operatorname{rank}\left\{(\mathbb{1}_{N_F}^{\top} \otimes \mathbf{H}[J_i, :])\right\} = 1$. Setting this upper bound to $N_s r$ gives the statement. □

C.4 Proof of Theorem 5.3

Proof. We omit the superscript l. It suffices to prove the case where $\tilde{\Phi}$ is square, $N_k = r$ and $N_F = N_s$. Then $\operatorname{rank}\{\tilde{\Phi}\} \in \mathbb{C}^{N_s r \times N_s r} = N_s r$ if and only if

$$\theta(\mathbf{H}) \triangleq \det \tilde{\Phi} = \det[\bar{\mathbf{H}} \odot \bar{\mathbf{B}}] \neq 0.$$

Now, $\theta(\mathbf{H})$ is a multivariate polynomial in the entries of $\mathbf{H}$ whose coefficents depend only on the entries of $\bar{\mathbf{B}}$. Thus $\theta(\mathbf{H})$ is either identically zero or its zero set is an affine algebraic set and thus a nowhere dense set of measure zero. It suffices to show $\theta(\mathbf{H}) \neq 0$ for a single choice of $\mathbf{H}$ (see, *e.g.*, [200–202] and references therein).

We can permute the rows of $\tilde{\Phi}$ such that the first N_k rows are indexed by J_1, the next N_k rows by J_2, and so on. In particular, there is a permutation matrix $\Pi \in \mathbb{C}^{N_k N_s \times N_k N_s}$ such that (*c.f.* (5.13))

$$\Pi\tilde{\Phi} = \begin{bmatrix} \mathbf{D}_1[J_1, J_1]\hat{\mathbf{B}}_1[J_1, :] & \ldots & \mathbf{D}_{N_F}[J_1, J_1]\hat{\mathbf{B}}_1[J_1, :] \\ \vdots & & \vdots \\ \mathbf{D}_1[J_1, J_1]\hat{\mathbf{B}}_{N_F}[J_1, :] & \ldots & \mathbf{D}_{N_F}[J_1, J_1]\hat{\mathbf{B}}_{N_F}[J_1, :] \\ \mathbf{D}_1[J_2, J_2]\hat{\mathbf{B}}_1[J_2, :] & \ldots & \mathbf{D}_{N_F}[J_2, J_2]\hat{\mathbf{B}}_1[J_2, :] \\ \vdots & \ddots & \vdots \\ \mathbf{D}_1[J_{N_F}, J_{N_F}]\hat{\mathbf{B}}_{N_F}[J_{N_F}, :] & \ldots & \mathbf{D}_{N_F}[J_{N_F}, J_{N_F}]\hat{\mathbf{B}}_{N_F}[J_{N_F}, :] \end{bmatrix} = \begin{bmatrix} \check{\mathbf{D}}_1^{J_1}\mathbf{C}_1 & \ldots & \check{\mathbf{D}}_{N_F}^{J_1}\mathbf{C}_1 \\ \check{\mathbf{D}}_1^{J_2}\mathbf{C}_2 & \ldots & \check{\mathbf{D}}_{N_F}^{J_2}\mathbf{C}_2 \\ \vdots & \ddots & \vdots \\ \check{\mathbf{D}}_1^{J_{N_F}}\mathbf{C}_{N_F} & \ldots & \check{\mathbf{D}}_{N_F}^{J_{N_F}}\mathbf{C}_{N_F} \end{bmatrix},$$

where, in an abuse of notation, we write $\check{\mathbf{D}}_l^J = (\mathbb{1}_{N_F}^{\top} \otimes \mathbf{D}_l[J, J])$.

Next, we specify our choice of $\mathbf{H}$. By assumption, $r = N_k = mN_F$ for some integer m. For each $i \in [N_s]$, we set $\mathbf{h}_i[J_i] = \mathbf{1}_m$ and set remaining coordinates are set to zero. With this construction, $\check{\mathbf{D}}_i^{J_j} = \mathbf{I}_m$ if $i = j$; otherwise, $\check{\mathbf{D}}_i^{J_j} = \mathbf{0}_m$. Now

$$\Pi\tilde{\Phi} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{0}_m & \ldots & \mathbf{0}_m \\ \mathbf{0}_m & \mathbf{C}_2 & \ldots & \mathbf{0}_m \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_m & \mathbf{0}_m & \ldots & \mathbf{C}_{N_F} \end{bmatrix},$$

that is, $\tilde{\Phi}$ is similar to a block diagonal matrix. As each block along the diagonal is full rank by assumption, $\tilde{\Phi}$ is full rank. □

References

[1] K. Jetter, G. Pfander, and G. Zimmermann, “The crest factor for trigonometric polynomials. Part I: Approximation theoretical estimates,” *Rev. Anal. Numér. Théor. Approx.*, vol. 30, no. 2, pp. 179–195, 2001.

[2] B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, ser. Signals and Communication Technology. Springer International Publishing, 2017.

[3] K. G. Murty and S. N. Kabadi, “Some NP-complete problems in quadratic and nonlinear programming,” *Mathematical Programming*, vol. 39, no. 2, pp. 117–129, 1987.

[4] P. A. Parrilo, “Semidefinite programming relaxations for semialgebraic problems,” *Mathematical Programming*, vol. 96, no. 2, pp. 293–320, 2003.

[5] P. A. Parrilo and B. Sturmfels, “Minimizing polynomial functions,” *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pp. 83–100, 2001.

[6] M. Vetterli, “A theory of multirate filter banks,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 3, pp. 356–372, 1987.

[7] P. Vaidyanathan, *Multirate Systems and Filter Banks.* Prentice Hall, 1992.

[8] M. N. Do, “Multidimensional filter banks and multiscale geometric representations,” *Foundations and Trends in Signal Processing*, vol. 5, no. 3, pp. 157–164, 2012.

[9] S. Venkataraman and B. Levy, “State space representations of 2-D FIR lossless transfer matrices,” *IEEE Trans. Circuits Syst. II*, vol. 41, no. 2, pp. 117–132, 1994.

[10] J. Zhou, M. Do, and J. Kovacevic, “Multidimensional orthogonal filter bank characterization and design using the Cayley transform,” *IEEE Trans. Image Process.*, vol. 14, no. 6, pp. 760–769, 2005.

[11] F. Delgosha and F. Fekri, “Results on the factorization of multidimensional matrices for paraunitary filterbanks over the complex field,” *IEEE Trans. Signal Process.*, vol. 52, no. 5, pp. 1289–1303, 2004.

[12] Z. Cvetkovic and M. Vetterli, “Oversampled filter banks,” *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1245–1255, May 1998.

[13] B. S. Albrecht Böttcher, *Introduction to Large Truncated Toeplitz Matrices.* Springer New York, 1999.

[14] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2005.

[15] T. F. Chan and J. A. Olkin, "Circulant preconditioners for Toeplitz-block matrices," *Numerical Algorithms*, vol. 6, no. 1, pp. 89–101, 1994.

[16] R. H. Chan, J. G. Nagy, and R. J. Plemmons, "Circulant preconditioned Toeplitz least squares iterations," *SIAM Journal on Matrix Analysis and Applications*, vol. 15, no. 1, pp. 80–97, Jan. 1994.

[17] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophysical Journal International*, vol. 33, no. 3, pp. 347–366, 1973.

[18] T. Laudadio, N. Mastronardi, and M. V. Barel, "Computing a lower bound of the smallest eigenvalue of a symmetric positive-definite Toeplitz matrix," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4726–4731, 2008.

[19] A. Zygmund, *Trigonometric Series.* Cambridge University Press, 2005.

[20] T. Sørevik and M. A. Nome, "Trigonometric interpolation on lattice grids," *BIT Numerical Mathematics*, vol. 56, no. 1, pp. 341–356, 2015.

[21] H. Ehlich and K. Zeller, "Schwankung von polynomen zwischen gitterpunkten," *Mathematische Zeitschrift*, vol. 86, no. 1, pp. 41–44, 1964.

[22] G. Wunder and H. Boche, "Peak magnitude of oversampled trigonometric polynomials," *Frequenz*, vol. 56, no. 5-6, pp. 102–109, 2002.

[23] J. G. Proakis and D. K. Manolakis, *Digital Signal Processing*, 4th ed. Prentice Hall, 2006.

[24] R. S. Elias M. Stein, *Fourier Analysis: An Introduction.* Princeton University Press, 2003.

[25] B. Şicleru and B. Dumitrescu, "POS3POLY—A MATLAB preprocessor for optimization with positive polynomials," *Optim. Eng.*, vol. 14, no. 2, pp. 251–273, 2013.

[26] Y.-P. Lin and P. P. Vaidyanathan, "Theory and design of two-dimensional filter banks: A review," *Multidimensional Systems and Signal Processing*, vol. 7, no. 3-4, pp. 263–330, 1996.

[27] H. Bolcskei, F. Hlawatsch, and H. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans. Signal Process.*, vol. 46, no. 12, pp. 3256–3268, 1998.

[28] O. Christensen, *An Introduction to Frames and Riesz Bases.* Birkhäuser, 2003.

[29] G. Strang and T. Nguyen, *Wavelets and Filter Banks.* Wellesley College, 1996.

[30] E. J. Candes and D. L. Donoho, “New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities,” *Commun. Pure Appl. Math.*, vol. 57, no. 2, pp. 219–266, 2004.

[31] E. Candès, L. Demanet, D. Donoho, and L. Ying, “Fast discrete curvelet transforms,” *Multiscale Modeling & Simulation*, vol. 5, no. 3, pp. 861–899, 2006.

[32] W.-S. Lu, A. Antoniou, and H. Xu, “A direct method for the design of 2-D nonseparable filter banks,” *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 45, no. 8, pp. 1146–1150, 1998.

[33] Y. Chen, M. D. Adams, and W.-S. Lu, “Design of optimal quincunx filter banks for image coding via sequential quadratic programming,” in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, pp. 897–900.

[34] L. Pfister and Y. Bresler, “Learning filter bank sparsifying transforms,” *IEEE Transactions on Signal Processing*, vol. 67, no. 2, pp. 504–519, 2019.

[35] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, `arXiv:1412.6980 [cs.LG]`, 2014.

[36] H. Bolcskei, “A necessary and sufficient condition for dual Weyl-Heisenberg frames to be compactly supported,” *The Journal of Fourier Analysis and Applications*, vol. 5, no. 5, pp. 409–419, 1999.

[37] B. Sharif and Y. Bresler, “Generic feasibility of perfect reconstruction with short FIR filters in multichannel systems,” *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 5814–5829, 2011.

[38] T. Strohmer, *Finite-and Infinite-Dimensional Models for Oversampled Filter Banks*. Boston, MA: Birkhäuser Boston, 2001, pp. 293–315.

[39] J. Zhou and M. N. Do, “Multidimensional oversampled filter banks,” in *Wavelets XI*, Aug. 2005.

[40] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, 2006.

[41] S. Ravishankar and Y. Bresler, “Learning sparsifying transforms,” *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1072–1086, 2013.

[42] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–45, Dec. 2006.

[43] S. Ravishankar and Y. Bresler, “Closed-form solutions within sparsifying transform learning,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[44] S. Ravishankar and Y. Bresler, “Learning doubly sparse transforms for images,” *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4598–4612, 2013.

[45] S. Ravishankar and Y. Bresler, “Learning overcomplete sparsifying transforms for signal processing,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3088–3092.

[46] B. Wen, S. Ravishankar, and Y. Bresler, “Structured overcomplete sparsifying transform learning with convergence guarantees and applications,” *International Journal of Computer Vision*, Oct. 2014.

[47] L. Pfister and Y. Bresler, “Model-based iterative tomographic reconstruction with adaptive sparsifying transforms,” in *Proc. SPIE Computational Imaging XII*, C. A. Bouman and K. D. Sauer, Eds. SPIE, Mar. 2014.

[48] S. Ravishankar and Y. Bresler, “Sparsifying transform learning for compressed sensing MRI,” in *International Symposium on Biomedical Imaging*, 2013.

[49] L. Pfister and Y. Bresler, “Tomographic reconstruction with adaptive sparsifying transforms,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 6914–6918.

[50] L. Pfister and Y. Bresler, “Adaptive sparsifying transforms for iterative tomographic reconstruction,” in *International Conference on Image Formation in X-Ray Computed Tomography*, 2014, pp. 107 – 110.

[51] R. Rubinstein, T. Peleg, and M. Elad, “Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model,” *IEEE Trans. Signal Process.*, vol. 61, no. 3, pp. 661–677, Feb. 2013.

[52] M. Yaghoobi, S. Nam, R. Gribonval, M. E. Davies et al., “Analysis operator learning for overcomplete cosparse representations,” in *European Signal Processing Conference (EUSIPCO’11)*, 2011.

[53] M. Yaghoobi, S. Nam, R. Gribonval, and M. E. Davies, “Noise aware analysis operator learning for approximately cosparse signals,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 5409–5412.

[54] M. Yaghoobi, S. Nam, R. Gribonval, and M. E. Davies, “Constrained overcomplete analysis operator learning for cosparse signal modelling,” *IEEE Trans. Signal Process.*, vol. 61, no. 9, pp. 2341–2355, May 2013.

[55] S. Hawe, M. Kleinsteuber, and K. Diepold, “Analysis operator learning and its application to image reconstruction,” *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2138–2150, June 2013.

[56] S. Roth and M. J. Black, “Fields of experts,” *International Journal of Computer Vision*, vol. 82, no. 2, pp. 205–229, 2009.

[57] Y. Chen, R. Ranftl, and T. Pock, “Insights into analysis operator learning: From patch-based sparse models to higher order MRFs,” *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1060–1072, Mar. 2014.

[58] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1256–1272, 2016.

[59] J.-F. Cai, H. Ji, Z. Shen, and G.-B. Ye, "Data-driven tight frame construction and image denoising," *Applied and Computational Harmonic Analysis*, vol. 37, no. 1, pp. 89–105, 2014.

[60] M. A. T. Figueiredo, "Synthesis versus analysis in patch-based image priors," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 1338–1342.

[61] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2528–2535.

[62] V. Papyan, J. Sulam, and M. Elad, "Working locally thinking globally: Theoretical guarantees for convolutional sparse coding," *IEEE Transactions on Signal Processing*, vol. 65, no. 21, pp. 5687–5701, 2017.

[63] B. Wohlberg, "Efficient algorithms for convolutional sparse representations," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 301–315, 2016.

[64] C. Garcia-Cardona and B. Wohlberg, "Convolutional dictionary learning: a comparative review and new algorithms," *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 366–381, 2018.

[65] S. Muramatsu, "Structured dictionary learning with 2-D non-separable oversampled lapped transform," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[66] S. Muramatsu, M. Ishii, and Z. Chen, "Efficient parameter optimization for example-based design of nonseparable oversampled lapped transform," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3618–3622.

[67] S. Muramatsu, K. Furuya, and N. Yuki, "Multidimensional nonseparable oversampled lapped transforms: Theory and design," *IEEE Transactions on Signal Processing*, vol. 65, no. 5, pp. 1251–1264, 2017.

[68] G. Peyré and J. M. Fadili, "Learning analysis sparsity priors," in *The 9th International Conference on Sampling Theory and Applications (SampTA)*, 2011.

[69] Y. Chen, T. Pock, and H. Bischof, "Learning ℓ_1-based analysis and synthesis sparsity priors using bi-level optimization," in *Workshop on Analysis Operator Learning vs Dictionary Learning, NIPS 2012*, 2012.

[70] J. K. Martin Vetterli, *Wavelets and Subband Coding.* Prentice-Hall, 1995.

[71] H. S. Malvar, *Signal Processing with Lapped Transforms.* Artech Print on Demand, 1992.

[72] A. D. Poularikas, Ed., *Transforms and Applications Handbook.* CRC Press, 2010.

[73] V. Goyal, M. Vetterli, and N. Thao, "Quantized overcomplete expansions in R^N: analysis, synthesis, and algorithms," *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 16–31, 1998.

[74] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising," in *Wavelets and Statistics*, ser. Lecture Notes in Statistics, A. Antoniadis and G. Oppenheim, Eds. Springer-Verlag, 1995, pp. 125–150.

[75] H.-Y. Gao and A. G. Bruce, "Waveshrink with firm shrinkage," *Statistica Sinica*, pp. 855–874, 1997.

[76] R. Chartrand, "Shrinkage mappings and their induced penalty functions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1026–1029.

[77] R. Chartrand, "Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009, pp. 262–265.

[78] U. Schmidt and S. Roth, "Shrinkage fields for effective image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2774–2781.

[79] L. Pfister and Y. Bresler, "Automatic parameter tuning for image denoising with learned sparsifying transforms," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[80] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proceedings of the 21th International Conference on Artificial Neural Networks.* Berlin, Heidelberg: Springer-Verlag, 2011, pp. 52–59.

[81] M. Tsatsanis and G. Giannakis, "Principal component filter banks for optimal multiresolution analysis," *IEEE Trans. Signal Process.*, vol. 43, no. 8, pp. 1766–1777, 1995.

[82] P. Moulin and M. Mihcak, "Theory and design of signal-adapted FIR paraunitary filter banks," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 920–929, 1998.

[83] B. Xuan and R. Bamberger, "Multi-dimensional, paraunitary principal component filter banks," in *1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 1995, pp. 1488–1491.

[84] B. Xuan and R. Bamberger, "FIR principal component filter banks," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 930–940, 1998.

[85] M. A. Unser, "Extension of the Karhunen-Loeve transform for wavelets and perfect reconstruction filterbanks," in *Mathematical Imaging: Wavelet Applications in Signal and Image Processing*, Nov. 1993, pp. 45–56.

[86] L. Pfister and Y. Bresler, "Learning sparsifying filter banks," in *Proc. SPIE Wavelets & Sparsity XVI*, vol. 9597. SPIE, Aug. 2015.

[87] A. Klöckner, N. Pinto, Y. Lee, B. Catanzaro, P. Ivanov, and A. Fasih, "PyCUDA and PyOpenCL: A scripting-based approach to GPU run-time code generation," *Parallel Computing*, vol. 38, no. 3, pp. 157–174, 2012.

[88] L. E. Givon, T. Unterthiner, N. B. Erichson, D. W. Chiang, E. Larson, L. Pfister, S. Dieleman, G. R. Lee, S. van der Walt, B. Menn, T. M. Moldovan, F. Bastien, X. Shi, J. Schlüter, B. Thomas, C. Capdevila, A. Rubinsteyn, M. M. Forbes, J. Frelinger, T. Klein, B. Merry, N. Merill, L. Pastewka, L. Y. Liu, S. Clarkson, M. Rader, S. Taylor, A. Bergeron, N. H. Ukani, F. Wang, W.-K. Lee, and Y. Zhou, "scikit-cuda 0.5.3: a Python interface to GPU-powered libraries," May 2019.

[89] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[90] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, 2007.

[91] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 2862–2869.

[92] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *2011 IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 479–486.

[93] U. Schmidt, Q. Gao, and S. Roth, "A generative perspective on MRFs in low-level vision," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1751–1758.

[94] B. Wen, Y. Li, and Y. Bresler, "When sparsity meets low-rankness: Transform learning with non-local low-rank constraint for image restoration," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 2297–2301.

[95] B. Wen, Y. Li, and Y. Bresler, "The power of complementary regularizers: Image recovery via transform learning and low-rank modeling," *CoRR*, `arXiv:1808.01316 [cs.CV]`, 2018.

[96] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2, July 2001, pp. 416–423.

[97] B. Saleh, *Fundamentals of Photonics*. Hoboken, N.J: Wiley-Interscience, 2007.

[98] M. Born and E. Wolf, *Principles of Optics*, 7th ed. Cambridge University Press, 1999.

[99] T. S. Ralston, D. L. Marks, P. S. Carney, and S. A. Boppart, “Inverse scattering for optical coherence tomography,” *JOSA A*, vol. 23, no. 5, pp. 1027–1037, 2006.

[100] B. J. Davis, S. C. Schlachter, D. L. Marks, T. S. Ralston, S. A. Boppart, and P. S. Carney, “Nonparaxial vector-field modeling of optical coherence tomography and interferometric synthetic aperture microscopy,” *JOSA A*, vol. 24, no. 9, pp. 2527–2542, 2007.

[101] B. J. Davis, D. L. Marks, T. S. Ralston, P. S. Carney, and S. A. Boppart, “Interferometric synthetic aperture microscopy: Computed imaging for scanned coherent microscopy,” *Sensors*, vol. 8, no. 6, pp. 3903–3931, June 2008.

[102] T. S. Ralston, D. L. Marks, S. A. Boppart, and P. S. Carney, “Inverse scattering for high-resolution interferometric microscopy,” *Opt. Lett.*, vol. 31, no. 24, pp. 3585–3587, 2006.

[103] A. J. Devaney, “Nonuniqueness in the inverse scattering problem,” *Journal of Mathematical Physics*, vol. 19, no. 7, pp. 1526–1531, 1978.

[104] A. F. Fercher, W. Drexler, C. K. Hitzenberger, and T. Lasser, “Optical coherence tomography-principles and applications,” *Reports on Progress in Physics*, vol. 66, no. 2, pp. 239–303, 2003.

[105] R. Wu and M. N. Toksöz, “Diffraction tomography and multisource holography applied to seismic imaging,” *GEOPHYSICS*, vol. 52, no. 1, pp. 11–25, 1987.

[106] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*. Society of Industrial and Applied Mathematics, 2001.

[107] R. Kress, *Linear Integral Equations*, ser. Applied Mathematical Sciences. Springer New York, 2014.

[108] W. C. Chew, M. S. Tong, and B. Hu, “Integral equation methods for electromagnetic and elastic waves,” *Synthesis Lectures on Computational Electromagnetics*, vol. 3, no. 1, pp. 1–241, 2008.

[109] B. Anthony J. Devaney, Northeastern University, *Mathematical Foundations of Imaging, Tomography and Wavefield Inversion*. Cambridge University Press, 2012.

[110] R. H. Stolt, “Migration by Fourier transform,” *GEOPHYSICS*, vol. 43, no. 1, pp. 23–48, 1978.

[111] C. Cafforio, C. Prati, and F. Rocca, “SAR data focusing using seismic migration techniques,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 27, no. 2, pp. 194–207, 1991.

[112] M. Soumekh, *Synthetic Aperture Radar Signal Processing with MATLAB Algorithms*. New York: J. Wiley, 1999.

[113] K. Mayer, R. Marklein, K. Langenberg, and T. Kreutter, “Three-dimensional imaging system based on Fourier transform synthetic aperture focusing technique,” *Ultrasonics*, vol. 28, no. 4, pp. 241–255, 1990.

[114] Z. Li, J. Wang, and Q. H. Liu, “Interpolation-free Stolt mapping for SAR imaging,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 5, pp. 926–929, 2014.

[115] F. Andersson, R. Moses, and F. Natterer, “Fast Fourier methods for synthetic aperture radar imaging,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 1, pp. 215–229, 2012.

[116] L. Demanet, M. Ferrara, N. Maxwell, J. Poulson, and L. Ying, “A butterfly algorithm for synthetic aperture radar imaging,” *SIAM Journal on Imaging Sciences*, vol. 5, no. 1, pp. 203–243, 2012.

[117] H. Barrett, *Foundations of Image Science*. Hoboken, NJ: Wiley-Interscience, 2004.

[118] M. Bertero, C. D. Mol, and E. R. Pike, “Linear inverse problems with discrete data. I. General formulation and singular system analysis,” *Inverse Problems*, vol. 1, no. 4, pp. 301–330, 1985.

[119] M. Bertero, C. D. Mol, and E. R. Pike, “Linear inverse problems with discrete data: II. Stability and regularisation,” *Inverse Problems*, vol. 4, no. 3, pp. 573–594, 1988.

[120] G. de Villiers and E. R. Pike, *The Limits of Resolution*. CRC Press, 2016.

[121] H. Landau, “Sampling, data transmission, and the Nyquist rate,” *Proceedings of the IEEE*, vol. 55, no. 10, pp. 1701–1706, 1967.

[122] M. Soumekh, “Band-limited interpolation from unevenly spaced sampled data,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 1, pp. 110–122, 1988.

[123] H. Erdogan and J. Fessler, “Monotonic algorithms for transmission tomography,” *IEEE Trans. Med. Imag.*, vol. 18, no. 9, pp. 801–14, Sep. 1999.

[124] S. Ramani and J. Fessler, “A splitting-based iterative algorithm for accelerated statistical x-ray CT reconstruction,” *IEEE Trans. Med. Imag.*, vol. 31, no. 3, pp. 677–688, Mar. 2012.

[125] K. Sauer and C. Bouman, “A local update strategy for iterative reconstruction from projections,” *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 534–548, 1993.

[126] J. Eckstein and D. Bertsekas, “On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators,” *Mathematical Programming*, vol. 55, no. 1-3, pp. 293–318, 1992.

[127] 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, no. 1, pp. 1–122, 2010.

[128] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, “An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems,” *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, 2011.

[129] Y. Ding, L. Caucci, and H. H. Barrett, “Null functions in three-dimensional imaging of alpha and beta particles,” *Scientific Reports*, vol. 7, no. 1, pp. 15 807–15 817, 2017.

[130] E. Clarkson and H. Barrett, “Bounds on null functions of linear digital imaging systems,” *Journal of the Optical Society of America A*, vol. 15, no. 5, pp. 1355–1360, 1998.

[131] A. K. Jha, H. H. Barrett, E. C. Frey, E. Clarkson, L. Caucci, and M. A. Kupinski, “Singular value decomposition for photon-processing nuclear imaging systems and applications for reconstruction and computing null functions,” *Physics in Medicine and Biology*, vol. 60, no. 18, pp. 7359–7385, 2015.

[132] F. Natterer, *The Mathematics of Computerized Tomography*. Philadelphia: Society for Industrial and Applied Mathematics, 2001.

[133] M. Anastasio, X. Pan, and E. Clarkson, “Comments on the filtered backprojection algorithm, range conditions, and the pseudoinverse solution,” *IEEE Trans. Med. Imag.*, vol. 20, no. 6, pp. 539–542, 2001.

[134] M. Bertero and E. Pike, “Resolution in diffraction-limited imaging, a singular value analysis,” *Optica Acta: International Journal of Optics*, vol. 29, no. 6, pp. 727–746, 1982.

[135] M. Bertero and P. Boccacci, “Computation of the singular system for a class of integral operators related to data inversion in confocal microscopy,” *Inverse Problems*, vol. 5, no. 6, pp. 935–957, 1989.

[136] A. Burvall, H. H. Barrett, C. Dainty, and K. J. Myers, “Singular-value decomposition for through-focus imaging systems,” *Journal of the Optical Society of America A*, vol. 23, no. 10, pp. 2440–2448, 2006.

[137] D. Slepian, “On bandwidth,” *Proceedings of the IEEE*, vol. 64, no. 3, pp. 292–300, 1976.

[138] D. Slepian, “Some asymptotic expansions for prolate spheroidal wave functions,” *Journal of Mathematics and Physics*, vol. 44, no. 1-4, pp. 99–140, 1965.

[139] D. Slepian, “Prolate spheroidal wave functions, Fourier analysis, and uncertainty-V: The discrete case,” *Bell System Technical Journal*, vol. 57, no. 5, pp. 1371–1430, 1978.

[140] D. Slepian, “Some comments on Fourier analysis, uncertainty and modeling,” *SIAM Review*, vol. 25, no. 3, pp. 379–393, 1983.

[141] D. Slepian and H. O. Pollak, “Prolate spheroidal wave functions, Fourier analysis and uncertainty - I,” *Bell System Technical Journal*, vol. 40, no. 1, pp. 43–63, 1961.

[142] A. Zetti, *Sturm-Liouville Theory*. Providence, Rhode Island: American Mathematical Society, 2010.

[143] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I: Functional Analysis*. New York: Academic Press, 1980.

[144] P. B. Bailey, W. N. Everitt, and A. Zettl, "The sleign2 Sturm-Liouville code," *ACM Trans. Math. Software*, vol. 27, no. 2, pp. 143–192, 2001.

[145] R. C. Brown, D. B. Hinton, and S. Schwabik, "Applications of a one-dimensional Sobolev inequality to eigenvalue problems," *Differential Integral Equations*, vol. 9, no. 3, pp. 481–498, 1996.

[146] Q. Kong, H. Wu, and A. Zettl, "Dependence of the Nth Sturm-Liouville eigenvalue on the problem," *Journal of Differential Equations*, vol. 156, no. 2, pp. 328–354, 1999.

[147] A. C. King, J. Billingham, and S. R. Otto, *Differential Equations: Linear, Nonlinear, Ordinary, Partial.* Cambridge University Press, 2003.

[148] "*NIST Digital Library of Mathematical Functions*," http://dlmf.nist.gov/, Release 1.0.23 of 2019-06-15, f. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

[149] A. Dutt and V. Rokhlin, "Fast Fourier transforms for nonequispaced data," *SIAM Journal on Scientific Computing*, vol. 14, no. 6, pp. 1368–1393, 1993.

[150] J. Fessler and B. Sutton, "Nonuniform fast Fourier transforms using min-max interpolation," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 560–574, 2003.

[151] G. Beylkin, "On the fast Fourier transform of functions with singularities," *Applied and Computational Harmonic Analysis*, vol. 2, no. 4, pp. 363–381, 1995.

[152] T. S. Ralston, D. L. Marks, P. S. Carney, and S. A. Boppart, "Interferometric synthetic aperture microscopy," *Nature Physics*, vol. 3, no. 2, pp. 129–134, 2007.

[153] Y. Xu, X. K. B. Chng, S. G. Adie, S. A. Boppart, and P. S. Carney, "Multifocal interferometric synthetic aperture microscopy," *Opt. Express*, vol. 22, no. 13, pp. 16 606–16 618, 2014.

[154] R. Bhargava, "Infrared spectroscopic imaging: The next generation," *Appl. Spectrosc.*, vol. 66, no. 10, pp. 1091–1120, Oct. 2012.

[155] B. J. Davis, P. S. Carney, and R. Bhargava, "Theory of midinfrared absorption microspectroscopy: I. Homogeneous samples," *Anal. Chem.*, vol. 82, no. 9, pp. 3474–3486, May 2010.

[156] B. J. Davis, P. S. Carney, and R. Bhargava, "Theory of midinfrared absorption microspectroscopy: II. Heterogeneous samples," *Anal. Chem.*, vol. 82, no. 9, pp. 3487–3499, May 2010.

[157] A. L. Oldenburg, C. Xu, and S. A. Boppart, "Spectroscopic optical coherence tomography and microscopy," *IEEE J. Sel. Topics Quantum Electron.*, vol. 13, no. 6, pp. 1629–1640, 2007.

[158] U. Morgner, W. Drexler, F. X. Kärtner, X. D. Li, C. Pitris, E. P. Ippen, and J. G. Fujimoto, "Spectroscopic optical coherence tomography," *Opt. Lett.*, vol. 25, no. 2, pp. 111–114, 2000.

[159] N. Bosschaart, T. G. van Leeuwen, M. C. G. Aalders, and D. J. Faber, "Quantitative comparison of analysis methods for spectroscopic optical coherence tomography," *Biomedical Optics Express*, vol. 4, no. 11, pp. 2570–2584, 2013.

[160] B. Deutsch, R. Reddy, D. Mayerich, R. Bhargava, and P. S. Carney, "Compositional prior information in computational infrared spectroscopic imaging," *J. Opt. Soc. Am. A*, vol. 32, no. 6, pp. 1126–1131, June 2015.

[161] Z. Pei Liang, "Spatiotemporal imaging with partially separable functions," in *2007 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, - 2007, pp. 988–991.

[162] H. Nguyen, J. Haldar, M. Do, and Z.-P. Liang, "Denoising of MR spectroscopic imaging data with spatial-spectral regularization," in *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2010, pp. 720–723.

[163] H. Nguyen, X. Peng, M. Do, and Z.-P. Liang, "Spatiotemporal denoising of MR spectroscopic imaging data by low-rank approximations," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2011, pp. 857–860.

[164] H. Nguyen, X. Peng, M. Do, and Z.-P. Liang, "Denoising MR spectroscopic imaging data with low-rank approximations," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 78–89, 2013.

[165] R. E. Alvarez and A. Macovski, "Energy-selective reconstructions in X-ray computerised tomography," *Physics in Medicine and Biology*, vol. 21, no. 5, pp. 733–744, 1976.

[166] J. Cammin, J. S. Iwanczyk, and K. Taguchi, "Spectral/photon-counting computed tomography," in *Emerging Imaging Technologies in Medicine*. CRC Press, 2012, pp. 40–57.

[167] G. Fowles, *Introduction to Modern Optics*. New York: Dover Publications, 1989.

[168] M. A. Anastasio, Q. Xu, and D. Shi, "Multispectral intensity diffraction tomography: Single material objects with variable densities," *Journal of the Optical Society of America A*, vol. 26, no. 2, pp. 403–412, 2009.

[169] H. J. Landau and H. O. Pollak, "Prolate spheroidal wave functions, Fourier analysis and uncertainty - II," *Bell System Technical Journal*, vol. 40, no. 1, pp. 65–84, 1961.

[170] H. J. Landau and H. O. Pollak, "Prolate spheroidal wave functions, Fourier analysis and uncertainty-III: The dimension of the space of essentially time- and band-limited signals," *Bell System Technical Journal*, vol. 41, no. 4, pp. 1295–1336, 1962.

[171] A. Jain and S. Ranganath, "Extrapolation algorithms for discrete signals with application in spectral estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 4, pp. 830–845, 1981.

[172] F. A. Grünbaum, "Eigenvectors of a Toeplitz matrix: Discrete version of the prolate spheroidal wave functions," *SIAM Journal on Algebraic Discrete Methods*, vol. 2, no. 2, pp. 136–141, 1981.

[173] Z. Zhu, S. Karnik, M. A. Davenport, J. Romberg, and M. B. Wakin, "The eigenvalue distribution of discrete periodic time-frequency limiting operators," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 95–99, 2018.

[174] E. Wolf and M. Nieto-Vesperinas, "Analyticity of the angular spectrum amplitude of scattered fields and some of its consequences," *Journal of the Optical Society of America A*, vol. 2, no. 6, pp. 886–890, 1985.

[175] N. D. Sidiropoulos and R. Bro, "On the uniqueness of multilinear decomposition of n-way arrays," *Journal of Chemometrics*, vol. 14, no. 3, pp. 229–239, 2000.

[176] S. Khanna and C. R. Murthy, "Corrections to 'On the restricted isometry of the columnwise Khatri-Rao product'," *IEEE Transactions on Signal Processing*, vol. 67, no. 9, pp. 2387–2388, 2019.

[177] S. Khanna and C. R. Murthy, "On the restricted isometry of the columnwise Khatri-Rao product," *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1170–1183, 2018.

[178] A. Fengler and P. Jung, "On the restricted isometry property of centered self Khatri-Rao products," *CoRR*, `arXiv:1905.09245 [cs.IT]`, 2019.

[179] E. S. Allman, C. Matias, and J. A. Rhodes, "Identifiability of parameters in latent structure models with many observed variables," *The Annals of Statistics*, vol. 37, no. 6A, pp. 3099–3132, 2009.

[180] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan, "Smoothed analysis of tensor decompositions," in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing - STOC '14*, - 2014, pp. 594–603.

[181] A. E. Gamal, N. Naderializadeh, and A. S. Avestimehr, "When does an ensemble of matrices with randomly scaled rows lose rank?" in *2015 IEEE International Symposium on Information Theory (ISIT)*, 6 2015, pp. 1502–1506.

[182] Y. C. Eldar and H. Bolcskei, "Block-sparsity: Coherence and efficient recovery," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 2885–2888.

[183] Y. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042–3054, 2010.

[184] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[185] E. Elhamifar and R. Vidal, "Block-sparse recovery via convex optimization," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4094–4107, 2012.

[186] J. W. Goodman, *Statistical Optics (Wiley Series in Pure and Applied Optics)*. Wiley, 2015.

[187] R. Coifman, V. Rokhlin, and S. Wandzura, “The fast multipole method for the wave equation: A pedestrian prescription,” *IEEE Antennas and Propagation Magazine*, vol. 35, no. 3, pp. 7–12, 1993.

[188] L. L. Meng, M. Hidayetoglu, T. Xia, W. E. I. Sha, L. J. Jiang, and W. C. Chew, “A wideband 2-d fast multipole algorithm with a novel diagonalization form,” *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 12, pp. 7477–7482, 2018.

[189] M. Hidayetoglu, C. Pearson, I. E. Hajj, L. Gurel, W. C. Chew, and W.-M. Hwu, “A fast and massively-parallel inverse solver for multiple-scattering tomographic image reconstruction,” in *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 5 2018, pp. 64–74.

[190] B. Chen and J. J. Stamnes, “Validity of diffraction tomography based on the first Born and the first Rytov approximations,” *Appl. Opt.*, vol. 37, no. 14, pp. 2996–3006, 1998.

[191] M. Slaney, A. Kak, and L. Larsen, “Limitations of imaging with first-order diffraction tomography,” *IEEE Trans. Microw. Theory Techn.*, vol. 32, no. 8, pp. 860–874, 1984.

[192] M. A. Anastasio, D. Shi, and G. Gbur, “Multispectral intensity diffraction tomography reconstruction theory: Quasi-nondispersive objects,” *Journal of the Optical Society of America A*, vol. 23, no. 6, pp. 1359–1368, 2006.

[193] R. Horstmeyer, J. Chung, X. Ou, G. Zheng, and C. Yang, “Diffraction tomography with Fourier ptychography,” *Optica*, vol. 3, no. 8, pp. 827–835, 2016.

[194] T.-A. Pham, E. Soubies, A. Goy, J. Lim, F. Soulez, D. Psaltis, and M. Unser, “Versatile reconstruction framework for diffraction tomography with intensity measurements and multiple scattering,” *Optics Express*, vol. 26, no. 3, pp. 2749–2763, 2018.

[195] M. Hidayetoglu, W.-M. Hwu, and W. C. Chew, “Supercomputing for full-wave tomographic image reconstruction in near-real time,” in *2018 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting*, 7 2018, pp. 1841–1842.

[196] H.-Y. Liu, D. Liu, H. Mansour, P. T. Boufounos, L. Waller, and U. S. Kamilov, “SEAGLE: Sparsity-driven image reconstruction under multiple scattering,” *IEEE Transactions on Computational Imaging*, vol. 4, no. 1, pp. 73–86, 2018.

[197] Y. Sun, Z. Xia, and U. S. Kamilov, “Efficient and accurate inversion of multiple scattering with deep learning,” *Optics Express*, vol. 26, no. 11, pp. 14 678–14 688, 2018.

[198] E. Soubies, T.-A. Pham, and M. Unser, “Efficient inversion of multiple-scattering model for optical diffraction tomography,” *Optics Express*, vol. 25, no. 18, pp. 21 786–21 800, 2017.

[199] G. H. Golub and C. F. V. Loan, *Matrix Computations.* Johns Hopkins University Press, 2012.

[200] G. Harikumar and Y. Bresler, “FIR perfect signal reconstruction from multiple convolutions: Minimum deconvolver orders,” *IEEE Transactions on Signal Processing*, vol. 46, no. 1, pp. 215–218, 1998.

[201] G. Harikumar and Y. Bresler, “Perfect blind restoration of images blurred by multiple filters: Theory and efficient algorithms,” *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 202–219, 1999.

[202] T. Jiang, N. Sidiropoulos, and J. ten Berge, “Almost-sure identifiability of multidimensional harmonic retrieval,” *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 1849–1859, 2001.