Spectral Image Processing Theory and Methods:
Reconstruction, Target Detection, and
Fundamental Performance Bounds

by

Kalyani Krishnamurthy

Department of Electrical and Computer Engineering, Duke University

Date: __________________________

Approved:

Prof. Rebecca Willett, Advisor

Prof. David Brady

Prof. Lawrence Carin

Prof. Leslie Collins

Prof. Loren Nolte

Prof. Mauro Maggioni

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Electrical and Computer Engineering
in the Graduate School of Duke University

2011
ABSTRACT
(Signal Processing)
Spectral Image Processing Theory and Methods: Reconstruction, Target Detection, and Fundamental Performance Bounds
by
Kalyani Krishnamurthy
Department of Electrical and Computer Engineering, Duke University

Date: ______________________

Approved:

________________________
Prof. Rebecca Willett, Advisor

________________________
Prof. David Brady

________________________
Prof. Lawrence Carin

________________________
Prof. Leslie Collins

________________________
Prof. Loren Nolte

________________________
Prof. Mauro Maggioni

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering in the Graduate School of Duke University 2011
Copyright © 2011 by Kalyani Krishnamurthy
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence
Abstract

This dissertation presents methods and associated performance bounds for spectral image processing tasks such as reconstruction and target detection, which are useful in a variety of applications such as astronomical imaging, biomedical imaging and remote sensing. The key idea behind our spectral image processing methods is the fact that important information in a spectral image can often be captured by low-dimensional manifolds embedded in high-dimensional spectral data. Based on this key idea, our work focuses on the reconstruction of spectral images from photon-limited, and distorted observations. This dissertation presents a partition-based, maximum penalized likelihood method that recovers spectral images from noisy observations and enjoys several useful properties; namely, it (a) adapts to spatial and spectral smoothness of the underlying spectral image, (b) is computationally efficient, (c) is near-minimax optimal over an anisotropic Hölder-Besov function class, and (d) can be extended to inverse problem frameworks.

There are many applications where accurate localization of desired targets in a spectral image is more crucial than a complete reconstruction. Our work draws its inspiration from classical detection theory and compressed sensing to develop computationally efficient methods to detect targets from few projection measurements of each spectrum in the spectral image. Assuming the availability of a spectral dictionary of possible targets, the methods discussed in this work detect targets that either come from the spectral dictionary or otherwise. The theoretical performance
bounds offer insight on the performance of our detectors as a function of the number of measurements, signal-to-noise ratio, background contamination and properties of the spectral dictionary.

A related problem is that of level set estimation where the goal is to detect the regions in an image where the underlying intensity function exceeds a threshold. This dissertation studies the problem of accurately extracting the level set of a function from indirect projection measurements without reconstructing the underlying function. Our partition-based set estimation method extracts the level set of proxy observations constructed from such projection measurements. The theoretical analysis presented in this work illustrates how the projection matrix, proxy construction and signal strength of the underlying function affect the estimation performance.
Dedicated to my husband, Shivakumar Vasanth
Contents

Abstract iv
List of Figures xi
Acknowledgements xiv

1 Introduction 1
  1.1 Spectral image processing . . . . . . . . . . . . . . . . . . . . . . . 1
  1.2 Dissertation overview . . . . . . . . . . . . . . . . . . . . . . . . . . 2

2 Spectral intensity estimation 7
  2.1 Spectral Poisson intensity estimation . . . . . . . . . . . . . . . . . 8
    2.1.1 Problem formulation . . . . . . . . . . . . . . . . . . . . . . . 9
    2.1.2 Background review . . . . . . . . . . . . . . . . . . . . . . . . 11
  2.2 Multiscale spatiotemporal intensity estimation . . . . . . . . . . . . 13
    2.2.1 RDP estimators . . . . . . . . . . . . . . . . . . . . . . . . . . 16
    2.2.2 Estimation algorithm . . . . . . . . . . . . . . . . . . . . . . . 18
  2.3 Error analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
    2.3.1 Specification of Hölder-Besov function class . . . . . . . . . . 21
    2.3.2 Upper bounds on the risk function . . . . . . . . . . . . . . . . 23
    2.3.3 Lower bounds on the risk function . . . . . . . . . . . . . . . . 25
  2.4 Computational complexity . . . . . . . . . . . . . . . . . . . . . . . . 27
  2.5 Spectral image reconstruction . . . . . . . . . . . . . . . . . . . . . 28
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1.1 Problem formulation</td>
<td>74</td>
</tr>
<tr>
<td>4.2 Our contribution and relation with previous work</td>
<td>74</td>
</tr>
<tr>
<td>4.2.1 Previous work on level set estimation</td>
<td>75</td>
</tr>
<tr>
<td>4.2.2 Relationship with previous work on sparse support detection</td>
<td>76</td>
</tr>
<tr>
<td>4.2.3 Our contribution</td>
<td>77</td>
</tr>
<tr>
<td>4.3 Fast level set estimation from projection measurements</td>
<td>78</td>
</tr>
<tr>
<td>4.3.1 Performance improvement via DC offset subtraction</td>
<td>81</td>
</tr>
<tr>
<td>4.4 Relationship with wavelet-based methods</td>
<td>82</td>
</tr>
<tr>
<td>4.5 Experimental results</td>
<td>83</td>
</tr>
<tr>
<td>4.6 Conclusion</td>
<td>88</td>
</tr>
<tr>
<td>5 Conclusion</td>
<td>119</td>
</tr>
<tr>
<td>5.1 Contributions</td>
<td>119</td>
</tr>
<tr>
<td>5.2 Summary and future work</td>
<td>121</td>
</tr>
<tr>
<td>A Proofs of Theorems and Lemmas</td>
<td>123</td>
</tr>
<tr>
<td>A.1 Proofs of results form Chapter 2</td>
<td>123</td>
</tr>
<tr>
<td>A.1.1 Proof of Theorem 1</td>
<td>123</td>
</tr>
<tr>
<td>A.2 Proofs of results from Chapter 3</td>
<td>127</td>
</tr>
<tr>
<td>A.2.1 Proof of Theorem 3</td>
<td>127</td>
</tr>
<tr>
<td>A.2.2 Proof of Theorem 4</td>
<td>128</td>
</tr>
<tr>
<td>A.2.3 Proof of Corollary 5</td>
<td>131</td>
</tr>
<tr>
<td>A.2.4 Proof of Corollary 6</td>
<td>131</td>
</tr>
<tr>
<td>A.2.5 Proof of Theorem 7</td>
<td>133</td>
</tr>
<tr>
<td>A.3 Proofs of results from Chapter 4</td>
<td>134</td>
</tr>
<tr>
<td>A.3.1 Proof of Theorem 8</td>
<td>134</td>
</tr>
<tr>
<td>Bibliography</td>
<td>138</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Sample partition of a spatiotemporal data cube. ................. 15
2.2 Illustration of the anisotropic Hölder-Besov function class. .... 22
2.3 Spatiotemporal denoising results on the AVIRIS data (at spectral band 6). ................................................................. 31
2.4 Denoising of AVIRIS spectrum at location (31,160). .............. 33
2.5 Spatiotemporal reconstruction results on the AVIRIS data (at spectral band 64). ............................................................. 34
2.6 Denoising results on real Chandra data. .............................. 37
3.1 Illustration of the anomaly detection hypothesis tests. ............ 59
3.2 Target detection simulation results. .................................. 63
3.3 Anomaly detection simulation results in the low SNR case. ....... 70
3.4 Anomaly detection results corresponding to real AVIRIS data. ... 71
4.1 Level set experimental results corresponding to BioImage1. ....... 89
4.2 Snapshots of level set estimation simulation results corresponding to BioImage1 when $K = N/2$. ........................................ 90
4.3 Level set experimental results corresponding to BioImage2. ....... 91
4.4 Snapshots of level set estimation simulation results corresponding to BioImage2 when $K = N/2$. ........................................ 92
4.5 Level set experimental results corresponding to BioImage3. ....... 93
4.6 Snapshots of level set estimation simulation results corresponding to BioImage3 when $K = N/2$. ........................................ 94
4.7 Level set experimental results corresponding to BioImage4. ....... 95
4.28 Snapshots of level set estimation simulation results corresponding to HyMapImage1 when $K = N/2$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 116

4.29 Level set experimental results corresponding to HyMapImage2. . . . . 117

4.30 Snapshots of level set estimation simulation results corresponding to HyMapImage2 image when $K = N/2$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 118
Acknowledgements

I owe my sincere thanks to my advisor, Prof. Rebecca Willett, for being an integral part of my journey through the doctoral program. Her attention to detail and enthusiasm for fresh, challenging ideas have always been contagious. I have immensely benefited through my interactions with her and also by simply observing the way she conducts research and deals with day-to-day work pressures. She has been an excellent mentor and a guide by providing me with the space to discuss research ideas, career goals, and work-life balance. I would always be grateful to her for being patient, supportive, and having faith in me at times when my research was not going too well.

I would like to thank the reminder of my committee members: Prof. David Brady, for his continuous support and insightful classes that have tremendously benefited my research; Prof. Lawrence Carin, for challenging me with thought provoking research questions that helped me to evaluate my research ideas carefully; Prof. Leslie Collins, for her interest in my research and making me feel at ease by providing me a space in her student office in the first year of my PhD; Prof. Loren Nolte, for being very receptive to my research ideas and offering helpful feedback; and Prof. Mauro Maggioni, for being an excellent critic of my research and giving me honest feedback about my research and presentation skills.

Much of the research presented in this dissertation is due to the wonderful collaborators I have worked with. I am grateful to Prof. Maxim Raginsky, for being
a great collaborator, good friend and a guide. I appreciate the time he has spent carefully reading my papers, sitting through my practice talks and offering excellent feedback. Thanks to him for preventing me from making mistakes in my proofs by giving in to my temptations or frustrations. I should admit that I am envious of his photographic memory. Many thanks to Dr. Waheed Bajwa, who collaborated with me on the level set estimation research. I have learned, and will continue to learn a lot from his methodical and sincere approach to research. Thanks to Prof. Stephen Reynolds and Prof. Kazimierz Borkowski at NCSU for exploring my spatiotemporal denoising algorithm to extract information from real astronomical data.

I am grateful to my colleagues for making my life in graduate school enjoyable and pleasant. Zachary Harmany and Dr. Jorge Silva have been very good friends, often lending patient ears to my rants and ramblings. We have engaged in several research discussions which have greatly benefited me. Thank you both for your time. I would like to thank my former lab mate, Prof. Roummel Marcia (now a faculty at UC Merced), for demonstrating the art of making good, understandable presentations supported by clear explanations. I thank Duke University and my department for providing me with the right resources to conduct research. I am grateful to Prof. Manjunath, UCSB, for providing me an opportunity to work in his Vision research lab as a visiting scholar from Aug 2005 to Nov 2005, which, in a way served as my entry door to the Ph.D program.

Many thanks to my understanding and supportive family that has been the driving force of my professional life. I am grateful to my parents, Banumathy and Krishnamurthy, brothers, Ajay Krishnamurthy and Ashwin Krishnamurthy, and parents-in-law, Vasanthi and Vasanth, for their prayers from abroad and moral support. Finally, my sincere thanks to my husband, Shivakumar Vasanth, for providing unconditional love and affection, showing faith in me, surviving my depression episodes, and being great a companion for life. I dedicate this dissertation to him.
Copyrights and credits

Parts of this dissertation have been adapted from the following papers that I have coauthored, with permissions from the publishers, IEEE and SIAM:


This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of the Duke University’s products or services. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to pubs-permissions@ieee.org. By choosing to view this material, you agree to all provisions of the copyright laws protecting it.
Introduction

Advances in science and engineering have spurred the development of sophisticated hardware and software, which have increased our ability to collect information about scenes or objects of interest with improved accuracy. The transition in imaging technology, from historical black-and-white photography to digital color imaging and the more sophisticated spectral imaging, has tremendously benefited several applications such as remote sensing, astronomical imaging and biomedical imaging. This dissertation studies methods and associated theory of spectral image processing tasks such as reconstruction and target detection that are often important in different application areas.

1.1 Spectral image processing

Spectral images consist of a spatial map of intensities across a possibly large number of spectral bands or wavelengths. Because spectral signatures are unique for every chemical element, observing these spectra at a high spatial and spectral resolution provides information about the material properties of the scene with much more accuracy than is possible with conventional RGB images. Spectral data collected by
a measurement system are often contaminated by noise and distortion that result from the physics of the measurement system and imaging conditions. As a result, observed spectral data are subjected to processing methods that depend on (a) the nature of the scene of interest, (b) data collection system, (c) spatial and spectral resolution of the observed data, and (d) the information that one wants to extract from the data. Such methods are collectively referred to by the term *spectral image processing*.

### 1.2 Dissertation overview

Spectral images are often high-dimensional since they are acquired at high spatial and spectral resolution. For instance, spectral data collected by the NASA Airborne Visible InfraRed Imaging Spectrometer (AVIRIS) consists of 224 spectral channels, and $512 \times 512$ spatial locations per spectral channel [1]. Similarly, the size of spectral data collected by the NASA Hyperspectral Mapper (HyMap) is approximately $845 \times 512 \times 126$ [2]. Collecting and processing such high resolution data is expensive, time consuming and computationally intensive. *A central theme of this dissertation is to develop efficient and practical spectral image processing methods by exploiting the fact that important structures such as edges and singularities often reside in low-dimensional manifolds embedded in high-dimensional spectral data.*

In many applications, spectral images are used to understand the physics of an underlying phenomenon by carefully examining the observed spectral signatures [3–5], and detect spatial locations in the data that correspond to known or unknown target spectral signatures of interest [6–9]. In remote sensing of forest covers, for example, the spectral information helps to identify different types of vegetation and their chlorophyll content, which leads to a better understanding of the forest ecosystem [3]. In astronomy, spectral images are used to study solar flares [4] which are high-energy emissions from the Sun that have significant impact on terrestrial weather patterns.
In fluorescence microscopy, the spectra help to identify unwanted emissions due to background, autofluorescence and other contaminants [5]. These are some example applications where a faithful extraction of the input spectral image from noisy and distorted observations (spectral image reconstruction) is important to learn about the scene of interest.

Chapter 2 focuses on the problem of spectral image reconstruction from photon-limited, distorted observations. Photon-limitedness refers to a scenario in which the observed mean intensity per voxel is extremely low. Observations can be photon-limited because of (a) a weak source, (b) a large distance between the source and the detector, and (c) detector resolution limits. Furthermore, they are often corrupted by the distortion introduced by the measurement system and the imaging conditions. In this work, we model the photon-limited observations to be Poisson distributed, and present a multiscale, partition-based, maximum penalized likelihood estimation (MPLE) method to recover spectral intensities from such observations. In line with the theme of this dissertation, our partition-based method exploits the low-dimensional manifold embedding of the structures (edges and singularities) in the high-dimensional spectral data to recursively refine the spatiospectral estimates from coarser to finer levels. Such a recursive estimation process smoothens the data in regions where the spatiospectral intensity is varying smoothly and preserves important edge information where the intensity is inhomogeneous. In addition, our method also lets the partition sizes vary along the spatial and spectral coordinates so that the method adapts to varying degrees of spatial and spectral smoothness.

Existing multiresolution methods for one- and two-dimensional Poisson intensity estimation [10–19] would yield suboptimal results when extended to spectral intensities since they cannot adapt to varying degrees of spectral and spatial smoothness of spectral intensities. Methods that recover spectral intensities from Gaussian observations [20–23] can be extended to the Poisson spectral intensity estimation problem
by transforming the Poisson observations to Gaussian through some kind of variance stabilizing transforms (VST) [24–26]. However, [10, 27] show that such transformations are inaccurate when the observations are extremely photon-limited. In contrast, our Poisson intensity estimation method (a) avoids the use of VSTs, (b) adapts to varying degrees of spatial and spectral smoothness, (c) is computationally efficient and (d) is near-minimax optimal over the anisotropic Hölder-Besov function class (discussed in greater detail in Sec. 2.3.1). Experimental results in Sec. 2.6 demonstrate the effectiveness of our method even when the observed mean intensity is less than 0.1 photons per voxel. We also show in Sec. 2.5 that our denoising method can successfully be incorporated in an inverse problem framework to perform spectral image reconstruction.

While spectral image reconstruction is important in several example applications discussed above, there are instances where an accurate localization of the spatial locations in a spectral image that correspond to some targets of interest (spectral target detection) is more important than spectral image reconstruction. Examples include agricultural applications, where one is interested in correctly classifying weeds and crops [6]; remote sensing, where it is important to classify a forest cover into different known tree species [7]; surveillance applications, where one is interested in locating anomalous objects in a scene [8]; and biomedical imaging, where it is important to detect the locations of cancerous cells [9]. In such applications, one can significantly minimize the computation time by directly performing target detection on the observations without first performing spectral image reconstruction. Furthermore, recent advances in compressed sensing [28, 29] and the associated hardware [30–35] suggest that we might be able to minimize the required number of detectors by observing projection measurements at each spatial location, instead of collecting direct, high resolution measurements.

In Chapter 3, we present methods and fundamental performance bounds for spec-
tral target detection from few projection measurements (relative to the dimension of each spectrum) observed at each spatial location, without reconstructing the underlying spectral intensity from projection measurements. Specifically, we assume access to a spectral dictionary $\mathcal{D}$ of possible targets of interest and detect spectral targets that either belong to $\mathcal{D}$ (*Dictionary Spectrum Detection (DSD)*), or lie outside $\mathcal{D}$ (*Anomalous Spectrum Detection (ASD)*), from projection measurements. Target detection performance is often affected by (a) background contamination at every spatial location from neighboring objects, (b) sensor noise, (c) similarity among different targets of interest in $\mathcal{D}$, (d) signal strengths of targets, (e) design of the projection operator, and (f) the number of projection measurements. In our work, we cast the spectral target detection problem as a multiple hypothesis testing problem where the $i^{th}$ hypothesis test corresponds to the presence or absence of the target of interest at the $i^{th}$ spatial location. We develop a nearest-neighbor based detector to perform DSD and analyze its performance using the positive False Discovery Rate (pFDR) error measure [36]. This error measure is particularly attractive in our problem framework since it lets us analyze the errors in multiple hypothesis tests simultaneously. Our performance bounds on the pFDR illustrate how the detection performance varies as a function of factors (a) through (f) discussed above. In the context of the ASD problem, our nearest-neighbor based anomaly detector is designed to control the FDR below a desired level. FDR control is achieved by extending the $p$-value based procedure of Benjamini and Hochberg [37] to our problem framework. We develop our detection methods based on nearest-neighbor searches in order to exploit the availability of projection measurements that preserve distances among a finite collection of high-dimensional vectors [38–40]. Experimental results in Sec. 3.6 illustrate the tightness of our bounds.

Level set estimation is the problem of detecting specific regions where a function exceeds a threshold given noisy observations of that function. This problem can be
considered as a generalization of the spectral target detection problem studied in Chapter 3. Specifically, let $\tilde{g}_i = g_i + n_i \in \mathbb{R}^N$ represent observations of spectrum $g_i$ corrupted by noise $n_i$ at the $i^{th}$ spatial location. Suppose that one is interested in finding the locations where $g_i = g_i^{(T)}$ for some known target spectrum $g^{(T)}$. If we let $f_i = \langle g_i, g^{(T)} \rangle$, then the target detection problem is equivalent to finding the level set $S^* = \{i : f_i \geq \gamma\}$ of the discrete function $f = [f_1, f_2, \ldots, f_N]^T$ from noisy observations $\{\tilde{g}_i\}$.

Set estimation is important in several applications such as medical imaging, astronomy and digital elevation maps. In many such applications, however, one might observe linear projections of $f$ rather than observing $f$ directly. Some examples include tomographic imaging, where one measures tomographic projections of $f$, and compressed sensing, where one observes pseudo-random projections of $f$. In Chapter 4, we discuss methods for level set estimation from such projection measurements, without reconstructing the underlying function $f$. Our set estimation method draws its inspiration from sparse support estimation methods [41–43] in compressed sensing and partition-based level set estimation methods [44]. In particular, our method computes proxy observations from the projection measurements and estimates the level set of the proxy observations using a partition-based approach. Our theoretical analysis provide insights on the performance of our estimator with respect to the construction of proxy, choice of projection operator and the signal energy of $f$. Experimental results in Sec. 4.5 demonstrate the effectiveness of our method on realistic medical imaging simulation derived from a magnetic resonance angiography image of the brain.

We conclude this dissertation with a discussion on our key contributions and avenues for future work in Chapter 5.
This chapter studies photon-limited, spectral intensity estimation and proposes a spatially- and spectrally-adaptive, nonparametric method to estimate spectral intensities from Poisson observations. Specifically, our method searches over estimates defined over a family of recursive dyadic partitions in both the spatial and spectral domains, and finds the one that maximizes a penalized log likelihood criterion. The key feature of this approach is that the partition cells are \textit{anisotropic} across the spatial and spectral dimensions so that the method adapts to varying degrees of spatial and spectral smoothness, even when the respective degrees of smoothness are not known a priori. The proposed approach is based on the key insight that spatial boundaries and singularities exist in the same locations in every spectral band, even though the contrast or perceptibility of these features may be very low in some bands. The incorporation of this model into the reconstruction results in significant performance gains. Furthermore, for spectral intensities that belong to the anisotropic Hölder-Besov function class, the proposed approach is shown to be near-minimax optimal. The upper bounds on the risk function, which is the expected squared Hellinger distance between the true intensity and the estimate obtained us-
ing the proposed approach, matches the best possible lower bound up to a log factor for certain degrees of spatial and spectral smoothness. Experiments conducted on realistic data sets show that the proposed method can reconstruct the spatial and the spectral inhomogeneities very well even when the observations are extremely photon-limited (i.e., less than 0.1 photon per voxel).

2.1 Spectral Poisson intensity estimation

Spectral images can be thought of as a measurement of the spectrum of light transmitted or reflected from each spatial location in a scene. Although spectral imaging has the potential to be very useful in several applications, its use poses significant challenges. For instance, the number of photons hitting the detector might be very low (this is referred to as the photon-limited regime) because of either a weak source or a large distance between the source and the detector. Even if the source is not weak, the observed photon counts are binned by the imaging system according to their spatial location and wavelength. This means that increasing the spatial or spectral resolution of an imaging system decreases the number of photons detected in each spatiotemporal bin. Thus, low photon counts always impact the quality of spectral images, either as photon noise or resolution limits. In addition, the geometry of imaging systems may introduce further distortions of the scene (either intentionally or unintentionally). For instance, lenses in optical systems introduce spatial blurring artifacts, and innovative new spectral imaging systems [31, 45, 46] measure pseudo-random projections of the scene to exploit recent theoretical developments in compressed sensing [28, 29, 47]. These indirect measurements of a scene result in challenging photon-limited inverse problems.

In this work, we describe and analyze an approach to estimate a spectral intensity from either indirect or direct photon-limited measurements. In particular, we assume that the observations follow a spatially and spectrally inhomogeneous Pois-
son distribution, and make the following two contributions: (a) develop a spatially and spectrally adaptive multiscale algorithm for recovering spectral image intensities from Poisson observations, and (b) provide a theoretical characterization of the proposed approach and prove that it is near-minimax optimal over a broad class of spectral images.

The remainder of the work is organized as follows. In Sec. 2.1.1, we provide a mathematical formulation of the problem of interest. This is followed by a review of the existing literature on Poisson intensity estimation and inverse problems in Sec. 2.1.2. We describe the complexity-penalized likelihood estimation algorithm for denoising in Sec. 2.2 and discuss the near-optimality of this approach and its computational complexity in Sec. 2.3 and Sec. 2.4, respectively. In Sec. 2.5 we extend the denoising algorithm using an Expectation-Maximization (EM) framework to solve inverse problems. We demonstrate the effectiveness of the proposed approach in performing denoising and image reconstruction with numerical experiments in Sec. 2.6.1 and Sec. 2.6.2 respectively, followed by a discussion in Sec. 2.8. Proofs of the main theorems are relegated to the appendices.

2.1.1 Problem formulation

Let \( f \) be the true spatially- and spectrally- varying spectral image on \([0, 1]^3\), where the first two dimensions correspond to the spatial locations and the third dimension corresponds to the spectral bands. We let \( x_1, x_2, \) and \( \lambda \) denote the two spatial and one spectral arguments of \( f \), respectively. Let \( \mathbf{f} \) denote a sampled version of \( f \), where

\[
\mathbf{f}_{i_1,i_2,i_3} = \int_{i_1/N_1}^{(i_1+1)/N_1} \int_{i_2/N_2}^{(i_2+1)/N_2} \int_{i_3/M}^{(i_3+1)/M} f(x_1, x_2, \lambda) d\lambda dx_2 dx_1
\]

(2.1)

for \( i_1 = 0, 1, \ldots, N_1 - 1 \), \( i_2 = 0, 1, \ldots, N_2 - 1 \) and \( i_3 = 0, 1, \ldots, M - 1 \). At the detector, we observe noisy and distorted photon counts which are given by

\[
y \sim \text{Poisson}(A \mathbf{f}),
\]

(2.2)
where $A$ corresponds to the distortion induced by the measurement system and “Poisson” denotes the independent observations of an inhomogeneous Poisson process with intensity $A\mathbf{f}$ on a grid of size $N_x \times N_y \times M$. The methods described in this work can easily be generalized to the case $N_x \neq N_y$, but carrying this notation throughout our proof makes it less transparent. We thus assume that $N_x = N_y = N$ for simplicity of presentation. Let $n$ denote the total number of observed events:

$$n \triangleq \sum_{i_1,i_2,i_3} y_{i_1,i_2,i_3}.$$

Our goal is to estimate $\mathbf{f}$ from $y$ as accurately as possible by exploiting anisotropic correlations in the spectral and spatial dimensions. The inference methods and associated theoretical properties described below assume that $f$ is piecewise smooth, meaning that spatially it is composed of smooth surfaces separated by smooth boundaries in the two spatial dimensions (and hence can be modeled as a member of a piecewise Hölder function class [48]), and that each spectrum in $f$ varies smoothly as a function of wavelength except for a finite number of singularities and discontinuities (and hence can be modeled as a member of a Besov function class [49]). Such a function class can model several complex spectral intensities like the flare spectral intensities [4] and thus has wide applicability.

We thus assume that $f$ belongs to an anisotropic Hölder-Besov function class (formally defined in detail in Sec. 2.3) and that $0 < C_t \leq f \leq C_u < \infty$. As discussed in Sec. 2.3, this anisotropy will allow us to model spectral images that exhibit different degrees of smoothness and irregularity in spatial and spectral domains, and precludes using isotropic basis functions (such as 3d wavelets) without sacrificing performance.

In this work, we will use the conventional $\mathcal{O}$ notation to specify the computational complexity and use the symbols $\preceq$ and $\asymp$ to describe the error rates, where $f_n \preceq g_n$ means there exists some $C > 0$ such that the sequences $f_n$ and $g_n$ satisfy $f_n \leq Cg_n$ for $n$ sufficiently large and $f_n \asymp g_n$ means there exist some $C_1, C_2 > 0$ such that
\[ C_1 g_n \leq f_n \leq C_2 g_n \text{ for } n \text{ sufficiently large.} \]

### 2.1.2 Background review

Many researchers have studied multiresolution methods to perform intensity estimation from Poisson data because of the ability of those methods to capture the inhomogeneities in the data; see, \([12, 14, 17, 50–54]\). Preserving discontinuities is critical because they potentially convey important information about the signal or image under observation. The key challenge in Poisson intensity estimation problems is that the mean and the variance of the observed counts are the same. As a result, the conventional wavelet-based approaches like hard thresholding \([55]\) and soft thresholding \([56]\), originally designed to denoise Gaussian data, will yield suboptimal results when applied to Poisson data with low intensities.

Variance-stabilizing transforms (VSTs), such as the Anscombe transform \([24]\) and the Haar-Fisz \([25, 26]\) transform, are widely used to address this issue and to approximate the Poisson observations by Gaussian random variables \([52, 57]\). Jansen proposes a wavelet based Poisson estimation method based on data-adaptive VSTs and Bayesian priors on the stabilized coefficients \([58]\). However, as pointed out in \([10, 27]\), such approximations are inaccurate when the observed number of photons per pixel or voxel is very low and tend to oversmooth the resulting estimate. In a more recent work, Zhang et al. \([59]\) propose a multiscale variance-stabilizing transform (MSVST) which applies a VST to the empirical wavelet, ridgelet or curvelet transform coefficients. However, theoretical analysis of this approach is not available and it is not clear how to extend the MSVST to Poisson inverse problems.

Several authors have investigated signal and image estimation methods specifically designed for Poisson noise without the need for VSTs. For example, Kolaczyk proposes scale-dependent corrected Haar wavelet thresholds for Poisson data \([10, 11]\). Bayesian approaches offer an elegant way of incorporating prior knowledge into the
estimation process to improve the performance. Kolaczyk [12] and Timmermann and Nowak [13] propose a multiscale approach using the unnormalized Haar wavelet transform in conjunction with Bayesian methods to perform denoising of Poisson data. In [14,15], Nowak and Kolaczyk extend the multiscale and Bayesian approaches mentioned above to Poisson inverse problems, where they present a MAP estimation algorithm in an Expectation-Maximization (EM) framework to reconstruct two-dimensional intensities. While Bayesian methods are optimal when the prior distribution accurately reflects the true distribution underlying the phenomenon being observed (i.e. $f$), it is not clear how the performance of these approaches changes with inaccurate approximations of the true prior.

In their seminal paper [16], Kolaczyk and Nowak present a multiscale framework for likelihoods similar to the multiresolution analysis on wavelets and propose a denoising algorithm based on the complexity-penalized likelihood estimation (CPLE). Compared to the Bayesian methods discussed above, the CPLE algorithm has fewer tuning parameters (or, alternately, avoids complex MCMC methods required when replacing tuning parameters with Bayesian hyperparameters) and is also minimax optimal over a wide range of isotropic likelihood models. There are variants of the CPLE method depending upon the nature of the image or signal being denoised [17–19].

Though there is a rich literature on one- and two-dimensional Poisson intensity estimation problems, there are not many algorithms developed for photon-limited Poisson spectral intensity estimation, which is the focus of this work. Extending the multiresolution methods discussed above in the context of one- and two-dimensional signals and images to spectral intensities will, in many applications, yield suboptimal results because the spatial and the spectral content of a spectral intensity might exhibit different degrees of smoothness. Conventional wavelet based approaches (e.g., 3d wavelets), cannot optimally adapt in this setting.

Atkinson et al. [20] propose a wavelet-based method for estimating spectral in-
tensities from noisy Gaussian observations. The basic idea behind their approach is to decorrelate the spectral data along the spectral dimension using the discrete Fourier transform, denoise each spatial map of Fourier coefficients independently using 2d wavelet-based thresholding algorithms, and reform the images using the inverse Fourier transform. Manjón, Robles and Thacker [21] propose an extension of the non-local means [22] filter to spectral measurements for denoising spectral MR images from Gaussian data. Scheunders suggests denoising spectral images from Gaussian observations using the interband correlations in the data [23]. However, these spectral intensity estimation algorithms are all designed for Gaussian noise statistics, and the impact of Poisson noise on their performance is not well understood. Some researchers have also considered the related problem of compressing spectral images [60], that, like our work, often relies on a low-complexity representation of the spectral image.

Recent work [61,62] on marked Poisson processes suggests that the use of marks (i.e. spectral dimension of a spectral intensity) can improve the spectral estimation accuracy when the observations are in the photon-limited regime. In this work, we extend the approach of [61,62] to a broader setting and show that the proposed algorithm is near minimax optimal in a certain anisotropic Hölder-Besov function class. We also demonstrate its effectiveness through the experiments conducted on realistic data sets.

2.2 Multiscale spatiospectral intensity estimation

In this section, we assume that the measurements collected at the detector are noisy but not distorted, so that \( A \) in (2.2) is the identity matrix, and describe our approach and analysis in this special case. The extension of these ideas to inverse problems will be discussed later in the chapter. Thus, we let \( y \sim \text{Poisson}(f) \).

Our approach consists of finding the spectral image within a class of candidate
estimates which optimizes a penalized log likelihood function. The class of candidate estimates and the penalty term are chosen to yield an estimator which is both near-minimax optimal for a broad and realistic class of spectral images and easy to compute rapidly. Specifically, we search over a family of recursive dyadic partitions (RDPs) of $[0, 1]^3$, and for each partition consider the estimate formed by computing maximum likelihood model fits on each partition cell. We then choose the partition which gives the best fit to the data (in a log likelihood sense) and which has low complexity. This can formally be expressed as

$$\hat{f} \equiv \arg \min_{f \in \Gamma_{M,N}} \left\{ -\log p(y|f) + \text{pen}(f) \right\}, \quad (2.3)$$

where $\Gamma_{M,N}$ is the class of candidate estimates and $\text{pen}(\cdot)$ is a complexity penalization term which satisfies the Kraft inequality [63] given by $\sum_{f \in \Gamma_{M,N}} e^{-\text{pen}(f)} \leq 1$. The Kraft inequality plays an important role in our proof of an upper bound on the estimation error. Intuitively, the penalties can be thought of as negative log prior probabilities assigned to each candidate estimator, and the Kraft inequality ensures that the prior probabilities sum to something less than one.

The class of candidate estimators, $\Gamma_{M,N}$, corresponds to functions which are piecewise constant spatially and piecewise polynomial spectrally, where the breakpoints between the constant and polynomial pieces are constrained by a recursive dyadic partition (RDP). The role of the RDP framework is to allow more localized model fits in regions where the intensity is very inhomogeneous (such as near a boundary) and hence preserve that inhomogeneity, and yet use much less local model fits in regions where there is strong homogeneity. This encourages significant smoothing in homogenous regions and removes noisy artifacts without eliminating key features.

The penalty of the estimate $\hat{f}$, which will be specified shortly, is proportional to the sum of the polynomial order of the model fits across all the cells in the partition.
corresponding to \( \hat{f} \). It is thus a measure of the estimator complexity and helps to balance the bias and the variance of the estimator; a higher penalty value favors smoother estimate with lower variance and higher bias, and a lower penalty value encourages complex estimates that have high fidelity to the observed data and a high variance.

The proposed approach takes advantage of correlations in the spectral image both between different spectral bands and between neighboring pixels by (a) dividing the spatial domain into spatially homogeneous regions, and (b) computing an estimate of the spectrum in each spatial region by dividing it into spectrally homogeneous or smooth bands. A sample partition of this kind is displayed in Figure 2.2. This approach leverages the key fact that spatial features such as boundaries between different spatial structures are manifested in the same spatial locations at all spectral bands, even though such features may have low contrast or be difficult to detect in certain spectral bands. Intuitively, we use high-contrast spectral bands to infer important information about the locations of the boundaries of these structures even in low-contrast spectral bands. Estimation procedures that do not leverage this fact can be much more vulnerable to noise or oversmoothing.

![Spatial Dimension 2](Spatial Dimension 2)

![Spatial Dimension 1](Spatial Dimension 1)

![Spectral Dimension](Spectral Dimension)

**Figure 2.1:** Sample partition of a spatiotemporal data cube. The spatial partition is the same at each spectral band, ensuring that the spatial boundaries are at the same locations in every spectral band.
2.2.1 RDP estimators

We define a dyadic spatial partition $\mathcal{P}$ on $[0,1]^3$ to be a disjoint union of dyadic cuboids that covers $[0,1]^3$, where each cuboid is of the form $[k_1 2^{-j}, (k_1 + 1)2^{-j}] \times [k_2 2^{-j}, (k_2 + 1)2^{-j}] \times [0,1]$ for a scale parameter $j \in \{0, 1, \ldots, \log_2(N)\}$ and indices $k_1, k_2 \in \{0, 1, \ldots, 2^j - 1\}$. Similarly, each spectrum can be represented by an RDP defined using disjoint dyadic intervals on $[0,1]$. A spatiospectral partition $\mathcal{Q}$ on $[0,1]^3$ is thus a disjoint union of dyadic cuboids, where each cuboid of the form $[k_1 2^{-j}, (k_1 + 1)2^{-j}] \times [k_2 2^{-j}, (k_2 + 1)2^{-j}] \times [k_3 2^{-j'}, (k_3 + 1)2^{-j'}]$, for scale parameters $j \in \{0, 1, \ldots, \log_2(N)\}$, $j' \in \{0, 1, \ldots, \log_2(M)\}$ and indices $k_1, k_2 \in \{0, 1, \ldots, 2^j - 1\}$ and $k_3 \in \{0, 1, \ldots, 2^{j'} - 1\}$. The spatial and spectral RDPs can be represented in terms of quadtrees and binary trees, respectively, in which tree branches correspond to partition cells being subdivided into smaller partition cells with half the sidelength.

We refer the readers to [16,17] for more details on this.

We define the class of possible estimates $\Gamma_{M,N}$ as follows. Let $r$ be a positive integer which is an upper bound on the smoothness of the spectra. Consider the set of all 1d functions (i.e. spectra) $g' : [0,1] \to [C_{\ell}, C_{u}]$ that are piecewise polynomial where each polynomial piece is of order $r$. The polynomial pieces are defined on dyadic intervals corresponding to a 1d RDP and the polynomial coefficients are quantized to one of $\sqrt{n}$ levels. Each of these functions can then be clipped according to

$$
g_{\text{clipped}}(\lambda) = g'(\lambda) \cdot I_{\{g'(\lambda) > C_{\ell}\}}$$

so that the resulting function is positive. Let $\Gamma'_{M,N}$ denote the collection of all $g_{\text{clipped}}$ satisfying the above construction. (We focus on the above piecewise polynomial model for our analysis, but note that piecewise exponential models, such as those described in [64], are a simple extension of the above, where $\log(g')$ is a piecewise polynomial.)
The class of candidate spectral image estimates, $\Gamma_{M,N}$, can now be defined in terms of the class of candidate spectral estimates, $\Gamma'_M$. Consider all functions $g : [0, 1]^3 \rightarrow [C_L, C_u]$ with a corresponding RDP $\mathcal{P}$ such that

$$g(x_1, x_2, \lambda) = \sum_{c \in \mathcal{P}} g'_c(\lambda) I_{\{(x_1, x_2) \in c\}},$$

where $g'_c \in \Gamma'_{M,N}$ for all $c \in \mathcal{P}$, i.e., $g$ is spatially partitioned according to an RDP $\mathcal{P}$, and every spatial location in a given cell $c \in \mathcal{P}$ has a corresponding spectrum $g'_c$. Each element of the class $\Gamma_{M,N}$ corresponds to a sampled version of $g$, where the sampling is similar to sampling of $f$ described in (2.1).

The penalty, as outlined earlier, is a measure of the estimator complexity and is proportional to the number of cells in the spatiotemporal partition $Q$ of the estimate $\hat{f}$ and the order of the polynomial fits to the partition cells of $Q$. Specifically, we set the penalty of $\hat{f}$ to

$$\text{pen}(\hat{f}) = |Q| \left( \frac{10}{3} + \frac{r^2 \log_2 n}{2} \right) \log e 2,$$

(2.4)

where $n$ is the total number of observed photon counts. This penalty is simply proportional to the complexity of the estimate, i.e. the number of RDP cells times the number of polynomial coefficients in each cell. The other terms in the penalty expression are scaling factors which ensure than our penalty satisfies the Kraft inequality since our proof hinges on the application of the Li-Barron theorem (explained in detail in Appendix A.1.1), which presupposes that the Kraft inequality is satisfied. The origin of these terms is detailed in Appendix A.1.1. This particular choice of penalty leads to near-minimax estimators as discussed in Sec. 2.3 and yields an optimal balance of the bias and the variance terms of the estimates.
The optimization problem in (2.3) can be solved accurately and efficiently using the approach described in this section. An initial, complete recursive dyadic partition (RDP) of $[0, 1]^2$ is obtained by recursively partitioning $[0, 1]^2$ into cells with dyadic (power of two) sidelengths until the finest resolution (pixel-level) is reached. The optimal spatiotemporal partition of the data is found by ascending through every level of the quadtree (starting at one level above the leaves), finding the best spectral estimate to the data in each RDP cell at that level, and pruning quadtree branches based on the penalized likelihood criterion. We explain this in detail below.

Given a spatial partition $P$, the estimate $f(P)$ can be calculated by finding the “best” model fit to the observed spectrum over each cuboid in $P$. The spectral estimate for a given cuboid can be computed using 1d penalized-likelihood Poisson intensity estimation methods, such as those described in [17]. In particular, for each cuboid $c$ and for each spectral band we sum the observations. This yields Poisson observations, denoted $\tilde{y}^{(c)}$, of the aggregate spectrum $\tilde{f}_{i_3}^{(c)} = \sum_{(i_1, i_2) \in c} f_{i_1, i_2, i_3}$. We then estimate the intensity of each spectrum of this form. This can be accomplished by pruning an RDP representation of the spectrum; the spectral RDP can be represented using a binary tree, and the models fit to each terminal interval in the spectral RDP can be constants, polynomials or exponentials.

In each partition cell $c$, we compute the following penalized log likelihood

$$
\hat{f}^{(c)} = \arg\min_{\tilde{f}^{(c)} \in \Gamma_{M,N}} \left\{ L^{(c)} \right\}, \quad (2.5)
$$

where $L^{(c)} = -\log p(\tilde{y}^{(c)}|\tilde{f}^{(c)}) + \text{pen}(\tilde{f}^{(c)})$ and $p(\tilde{y}^{(c)}|\tilde{f}^{(c)})$ corresponds to the Poisson likelihood given by

$$
p(\tilde{y}^{(c)}|\tilde{f}^{(c)}) = \prod_{i_3=0}^{M-1} \frac{e^{-\tilde{f}^{(c)}_{i_3}} \left( \tilde{f}^{(c)}_{i_3} \right)^{\tilde{y}_{i_3}}}{\tilde{y}_{i_3}!}. \quad (2.6)
$$
We define $\text{pen}(\tilde{f}^{(c)})$ to be the penalty proportional to the number of terminal intervals in the pruned binary RDP; the penalties are discussed in detail in [16,17].

As the algorithm iterates over every level of the quadtree, it prunes the branches of the quadtree to find the partition $\mathcal{P}$ with the minimal sum of the $\{L^{(c)}\}$. The final spatiospectral estimate is then calculated by finding the partition $\hat{\mathcal{P}}$, which minimizes the total penalized likelihood function:

$$
\hat{\mathcal{P}} \equiv \arg \min_{\mathcal{P}} \left\{ \sum_{c \in \mathcal{P}} L^{(c)} \right\} \text{ and } \hat{f}_{i_1,i_2,i_3} \equiv \sum_{c \in \hat{\mathcal{P}}} \hat{f}^{(c)} I_{((i_1,i_2) \in c)}.
$$

(2.7)

In this work, we refer to this approach as the full spatiospectral denoising algorithm since at every level of quadtree pruning we perform both spatial and spectral smoothing. Each of the terminal intervals in the pruned RDP could correspond to a homogeneous or smoothly varying region of intensity. This endows our estimators with spatially and spectrally varying resolution to automatically increase the smoothing in very regular regions of the intensity and preserve detailed structures in less homogeneous regions.

This approach is similar to the image estimation method described in [16,65], with the key distinction that the proposed method forces the spatial RDP to be the same at every spectral band. This constraint ensures that the method preserves the spatial boundaries at the same locations in every spectral band, irrespective of the contrast differences among different spectral bands. In other words, when a tree branch is pruned, it means partition cells are merged in every spectral band simultaneously at the corresponding spatial location. This approach is effective because an outlier observation in one spatiospectral voxel may not be recognized as such when spectral bands are considered independently, but may be correctly pruned when the corresponding spectrum is very similar to a spatially neighboring spectrum.
2.3 Error analysis

In this section, we present a minimax upper bound on the risk of the full spatiotemporal denoising algorithm proposed in Sec. 2.2.2, and also derive minimax lower bounds on the risk of any spectral intensity estimation procedure to demonstrate the near-optimality of our approach over anisotropic H"older-Besov function class to be defined precisely below.

To facilitate the error analysis, let us reformulate the intensity estimation problem using a multinomial framework. The conditional distribution of a random variable $y \sim \text{Poisson}(f)$ with unit total intensity (i.e., the components of $f$ sum to one) given the observed number of photon counts $n$ is multinomial. Estimating the Poisson intensity can be broken into two components: (1) estimating the total intensity $I_f$ of the spectral image and (2) estimating the normalized spectral image $f/I_f$. The multinomial framework allows us to bound the error between $f/I_f$ and $\hat{f}/I_{\hat{f}}$. Unfortunately, the error in estimating $I_f$ can make it very difficult to bound the error between $f$ and $\hat{f}$ using the Poisson framework, even though the normalized error is primarily important to most end-users. Assuming that the true continuous-domain intensity $f$ integrates to unity, $I_f \equiv \int f = 1$, the components of $f$ will sum to unity as well. Consequently, we restrict each $\tilde{f} \in \Gamma_{M,N}$ to be positive and to sum to one. The observations $y$ are now assumed to follow a multinomial distribution with parameter vector $f$, that is, $y \sim \text{Multinomial}(f; n)$. We enumerate the voxels according to the lexicographic ordering and let $f_j$ be the component of $f$ corresponding to the $j^{\text{th}}$ voxel. The likelihood of observing $y$ given $f$ under this model is

$$p(y|f) = \frac{n!}{\prod_{i=1}^{N^2 M} f_i^y_i} f_1^{y_1} f_2^{y_2} \cdots f_{N^2 M}^{y_{N^2 M}}.$$  \hfill (2.8)
### 2.3.1 Specification of Hölder-Besov function class

We assume that the underlying density $f$ lies in an anisotropic Hölder-Besov function class $\mathcal{F}$ which we define as the space of functions on the unit cube $[0,1]^3$ that exhibit piecewise Hölder smoothness in the first two (spatial) dimensions and Besov smoothness in the third (spectral) dimension, as illustrated in Fig. 2.2. To describe such a Hölder-Besov function class we will draw upon the machinery of anisotropic smoothness classes [66]. We start by developing an appropriate notion of a modulus of smoothness following [49]. Consider a continuous function $f$ on $[0,1]^d$ for some $d \geq 1$. Given $r = 1, 2, \ldots$ and $h = (h_1, h_2, \ldots, h_d) \in \mathbb{R}^d$, let $\Delta^r_h f : \mathbb{R}^d \rightarrow \mathbb{R}$ be the $r$th difference of $f$ with step $h$:

$$
\Delta^r_h f(x_1, x_2, \ldots, x_d) \triangleq \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x_1 + kh_1, x_2 + kh_2, \ldots, x_d + kh_d).
$$

We will mostly deal with the case when $d$ is equal to 1, 2 or 3. Note that the mapping $f \mapsto \Delta^r_h f$ is linear and the function $\Delta^r_h f$ is supported on the set $A^{(d)}_{r,h} \triangleq \prod_{i=1}^{d} [0, \min\{0, 1 - rh_i\}]$. The corresponding moduli of smoothness are given by

$$
\omega_r(f, t)_p \triangleq \sup_{h \in \mathbb{R}^d, \|h\| \leq t} \|\Delta^r_h f\|_{L^p}
$$

where $r = 1, 2, \ldots$, $1 \leq p \leq \infty$, and let $\omega(f, t) \triangleq \omega_1(f, t)_\infty$.

We are interested in functions on $[0,1]^3$ that exhibit different kinds of smoothness in different coordinate directions. In particular, we consider functions that are Hölder in the first two directions and Besov in the third direction. We will say that a function $f : [0,1]^3 \rightarrow \mathbb{R}$ is spatially piecewise Hölder-(\(\alpha, \gamma\)) smooth if, for any fixed $\lambda \in [0,1]$ and all $(x_1, x_2) \in [0,1]^2$ we can write

$$
f(x_1, x_2, \lambda) = f_1(x_1, x_2, \lambda)I_{\{H(x_1) \geq x_2\}} + f_2(x_1, x_2, \lambda)I_{\{H(x_1) < x_2\}}
$$

(2.9)
where, for each $\lambda \in [0, 1]$, the surfaces $(x_1, x_2) \mapsto f_j(x_1, x_2, \lambda)$ for $j = 1, 2$ are Hölder-$\alpha$ for $\alpha \in (0, 1]$, so that
\[
|f_j(x_1, x_2, \lambda) - f_j(x'_1, x'_2, \lambda)| \leq C_\alpha ((x_1 - x'_1)^2 + (x_2 - x'_2)^2)^{\alpha/2} \quad (2.10)
\]
for any $(x_1, x_2), (x'_1, x'_2) \in [0, 1]^2$ [48]. Also, $H(x)$ is Hölder-$\gamma$ for $\gamma \in (0, 1]$ so that
\[
|H(x) - H(x')| \leq C_\gamma |x - x'|^\gamma \quad (2.11)
\]
In other words, for each $\lambda$, $f(x_1, x_2, \lambda)$ consists of two Hölder surfaces separated by a Hölder boundary. Further assume that for any fixed $(x_1, x_2) \in [0, 1]^2$, the one-dimensional function $\lambda \mapsto f(x_1, x_2, \lambda)$ is in the Besov space $B^\beta_\tau(L_p([0, 1]))$ where the smoothness parameter $\beta > 0$ and $1/p = \beta + 1/\tau$, where $L_p([0, 1])$ is the approximation space [49]. In this work, we fix $\tau = 2$.

![Figure 2.2](image)

**Figure 2.2**: Illustration of the anisotropic Hölder-Besov function class. (a) Spatial variations in $f(x_1, x_2, \lambda')$ for a fixed $\lambda'$. The intensity is piecewise smooth along the spatial coordinates with Hölder smooth surfaces separated by a Hölder smooth boundary. (b) Besov smooth spectrum corresponding to fixed spatial location $(x'_1, x'_2)$ that is piecewise smooth with finitely many discontinuities.

In terms of the moduli of smoothness, the surfaces $f_1$ and $f_2$ that enter into the definition in (2.9) belong to a family of functions $f : [0, 1]^3 \to \mathbb{R}$ such that for a given $1 \leq p, q < \infty$, $0 < \alpha \leq 1$, and $\beta > 0$,
\[
|f|^{(1, 2)}_{H_\alpha} \triangleq \sup_{0 \leq \lambda \leq 1} \sup_{t > 0} (t^{-\alpha} \omega(f(\cdot, \lambda), t)) < +\infty. \quad (2.12)
\]
\[ |f|_{B^\beta_q(L^p)}^{(3)} \triangleq \sup_{(x,y) \in [0,1]^2} \left( \int_0^1 (t^{-\beta} \omega_r(f(x,y,\cdot),t))^q \frac{dt}{t} \right)^{1/q} < +\infty, \tag{2.13} \]

where \( r = \lfloor \beta \rfloor + 1 \) [49]. The superscript \((1,2)\) on \(|f|_{H^{(1,2)}}\) in (2.12) refers to the fact that we measure the modulus of smoothness only along the first two spatial dimensions; the same goes for the superscript \((3)\) in (2.13). With these definitions, we can formalize the notion of different types of smoothness in different coordinate directions. In particular, the condition in (2.12) defines the functions \( f \) on \([0,1]^3\) that are uniformly Hölder in the first two coordinate directions. Similarly, the condition (2.13) defines the functions \( f \) on \([0,1]^3\) that are uniformly Besov in the third coordinate direction. Thus, \( f_1 \) and \( f_2 \) in (2.9) belong to the function class

\[ \mathcal{F}_{\alpha,\beta,p,q}(L) \triangleq \left\{ f : [0,1]^3 \to \mathbb{R} \left| \|f\|_{L^2} + |f|_{H^{(1,2)}_\alpha} + |f|_{B^\beta_q(L^p)} < L \right. \right\} \]

for some \( L > 0 \). We will see that although our estimate must be a member of the class \( \Gamma_{M,N} \), we can accurately estimate any function in the Hölder-Besov space.

### 2.3.2 Upper bounds on the risk function

The derivation of the upper bounds on the risk function follows the one in [17], but with some significant differences because of the fact that we consider anisotropic three-dimensional Hölder-Besov densities here, whereas [17] deals with one- and two-dimensional densities in the Besov and Hölder function spaces, respectively, and explicitly considers discrete-domain intensities.

We define the risk function between the true intensity \( f \) and its penalized log likelihood estimate \( \hat{f} \) (defined in (2.3)) as follows: \( R(f, \hat{f}) \equiv E \left[ \mathcal{H}^2(f, \hat{f}) \right] \) where

\[ \mathcal{H}^2(f, \hat{f}) = \sum_{j=1}^{N^2M} \left( \sqrt{f_j} - \sqrt{\hat{f}_j} \right)^2 \tag{2.14} \]
is the squared Hellinger distance between \(f\) and \(\hat{f}\). Here, the expectation is taken with respect to the observations.

**Theorem 1.** Let \(\mathcal{F}\) denote the class of functions of the form \((2.9)\). Suppose that the observation \(y \sim \text{Multinomial}(f; n)\), where \(f\) is obtained from an unknown intensity \(f \in \mathcal{F}\) with \(I_f = 1\) via integration sampling. Let \(\Gamma_{M,N}\) be a class of candidate estimators as detailed in Sec. 2.2, and let \(\hat{f}\) be the estimate obtained by the full spatiomatic denoising algorithm according to \((2.7)\). Assume that every \(\tilde{f} \in \Gamma_{M,N}\) satisfies the condition \(\tilde{f}_j \geq \frac{C}{N^2 M}\), for \(j = 1, \ldots, N^2 M\) for some \(C \in (0, 1)\), and that the penalty is given by \((2.4)\). Assume \(N\) and \(M\) are sufficiently large, so that

\[
M \log_2 M > C' \left( \frac{n \log_e n}{\log_2 n} \right)^{\frac{2\beta}{2\beta(\nu + 2 - \gamma) - \nu}}
\]

\[
N \log_2 M > C'' \left( \frac{n \log_e n}{\log_2 n} \right)^{\frac{2\beta}{2\beta(\nu + 2 - \gamma) - \nu}}
\]

where \(C'\) is a constant independent of \(n\), \(N\) and \(M\) and dependent on the smoothness parameters \(\alpha\), \(\beta\) and \(\gamma\). Assume that the polynomial fits in \((2.7)\) are of order \(r \geq \lceil \beta \rceil\). Then,

\[
R(f, \hat{f}) \preceq \left( \frac{n}{\log_2 M \log_e n} \right)^{-\frac{2\beta \nu}{2\beta(\nu + 2 - \gamma) - \nu}},
\]

\((2.15)\)

where \(\nu = \min(\alpha, \gamma)\).

The proof of Theorem 1 is given in Appendix A.1.1. As detailed in the proof, \(\log_2 M\) term arises from a combination of known approximation error bounds for free knot polynomials and our restriction of polynomial endpoints to boundaries in a recursive dyadic partition.

For a fixed \(n\), the lower bounds on \(N\) and \(M\) place a limit on the spatial and the spectral resolution of the observed data to ensure that the photon noise is dominant
over the errors due to sampling. In particular, if $M$ and $N$ are held fixed (and small), but the number of photons $n$ increases, at some point photon noise will be negligible relative to errors from sampling the spectral image to fit on the $N \times N \times M$ grid. In that case, the estimation error will not continue to decay with $n$ at the rate our bounds suggest. Since approximation errors due to sampling are not the focus of this work, we consider the case where $M$ and $N$ are relatively large.

2.3.3 Lower bounds on the risk function

In this section, we present lower bounds on the Hellinger risk when the unknown target intensity is a member of the Hölder-Besov class. To keep the technical details to a minimum, we consider the problem of estimating the actual continuous-domain intensity $f : [0, 1]^3 \to [0, 1]$, without discretization. Effectively, this corresponds to the case when the number of voxels $N^2 M \to \infty$, while the number of observed photon counts $n$ is held fixed. Since $\int_{[0,1]^3} f = 1$, this asymptotic scenario is equivalent to observing $n$ independent samples $z_1, \ldots, z_n$ from a probability density $f$ with support in the unit cube $[0,1]^3$, and the goal is to estimate $f$. To measure the quality of a candidate estimator $\hat{f}$, we adopt the Hellinger risk $R(f, \hat{f}) \doteq \mathbb{E}\left[\mathcal{H}^2(f, \hat{f})\right]$, where

$$
\mathcal{H}^2(f, \hat{f}) = \int_{[0,1]^3} \left(\sqrt{f(x_1, x_2, \lambda)} - \sqrt{\hat{f}(x_1, x_2, \lambda)}\right)^2 dx_1 dx_2 d\lambda
$$

and the expectation is taken w.r.t. $z_1, \ldots, z_n$. The squared Hellinger distance $\mathcal{H}^2(f, \hat{f})$ defined in (2.14) can be viewed as a discrete approximation of the above integral, and the limit $N, M \to \infty$ corresponds to taking increasingly fine approximations. Moreover, the set of all possible estimators of $f$ includes those that first bin the observations into voxels. Thus, the lower bounds on the risk of estimating the actual continuous-domain intensity $f$ also provide lower bounds for the discrete estimators of $f$. We are interested in lower bounds on the minimax risk
\( R_n(\mathcal{F}) \triangleq \inf_f \sup_{\hat{f} \in \mathcal{F}} R(f, \hat{f}) \), where the infimum is over all estimators \( \hat{f} \) based on \( n \) i.i.d. samples \( z_1, \ldots, z_n \) from \( f \).

Our derivation of the minimax lower bound relies on a powerful information-theoretic method of Yang and Barron [67]. The key idea behind the method of [67] is that the minimax rates of convergence for a wide variety of function classes can be determined from the global metric properties of a carefully chosen subset of the particular function class. These metric properties are encoded in the covering and the packing numbers of the class. Before stating the main result of [67], let us define the following. Given \( \epsilon > 0 \) and a function class \( \mathcal{G} \subset L_2([0,1]^d) \), a finite set \( S \subset \mathcal{G} \) is an \( \epsilon \)-packing set for \( \mathcal{G} \) if

\[
\min \left\{ ||g - g'||_{L_2([0,1]^d)} : g, g' \in S; g \neq g' \right\} \geq \epsilon. 
\]

Let \( M(\epsilon, \mathcal{G}) \) denote the cardinality of the maximal \( \epsilon \)-packing set for \( \mathcal{G} \) with respect to \( ||.||_{L_2([0,1]^d)} \); the Kolmogorov \( \epsilon \)-capacity of \( \mathcal{G} \) is defined by \( K_\epsilon(\mathcal{G}) \triangleq \log M(\epsilon, \mathcal{G}) \) [68].

One of the key results of [67] is that, if \( \mathcal{F} \supseteq \mathcal{G} \) is a class of density functions bounded above and below such that \( 0 < C_\ell \leq f \leq C_u < \infty \) for all \( f \in \mathcal{F} \), then we have the minimax lower bound \( R_n(\mathcal{F}) \succeq \epsilon_n^2 \), where \( \epsilon_n \) is the critical separation distance, determined from the Kolmogorov packing entropy by solving the equation

\[
K_{\epsilon_n}(\mathcal{G}) = n \epsilon_n^2. \tag{2.16}
\]

In this work, as indicated in Sec. 2.3, \( \mathcal{F} \) represents the anisotropic Hölder-Besov function class consisting of piecewise Holder surfaces separated by a Hölder boundary in the spatial dimensions and piecewise Besov spectra in the spectral dimension. Using the result from [67] referred to above, we arrive at the following lower bound.

**Theorem 2.** (Minimax lower bounds [69]) Let us assume that we observe \( n \) i.i.d. realizations drawn from a density \( f \in \mathcal{F} \) and that \( 0 < C_\ell \leq f \leq C_u < \infty \). Let \( \hat{f} \) be
any estimate of $f$ based on these $n$ realizations. Then the minimax lower bound is given by

$$
\mathbb{E} \left[ \mathcal{H}^2(f, \hat{f}) \right] \geq \max \left( n^{-2\gamma/(2\gamma+1)}, n^{-2\alpha\beta/(2\alpha\beta+2\beta+\alpha)} \right)
$$

(2.17)

for $\alpha \in (0, 1]$, $\gamma \in (0, 1]$ and $\beta > 0$.

The proof of Theorem 2 is given in [69]. For $\gamma = 1$, in particular, we have

$$
\nu = \min(\alpha, \gamma) = \alpha,
$$

and

$$
\mathbb{E} \left[ \mathcal{H}^2(f, \hat{f}) \right] \geq \max \left( n^{-2/3}, n^{-2\alpha\beta/(2\alpha\beta+2\beta+\alpha)} \right) \equiv n^{-2\alpha\beta/(2\alpha\beta+2\beta+\alpha)}.
$$

Thus the lower bound matches the upper bound given in (2.15) up to a log factor when $\gamma = 1$.

### 2.4 Computational complexity

Implementing the full spatiotemporal algorithm involves performing both spatial and spectral smoothing at every level of the tree. For a datacube of size $N \times N \times M$ the computational complexity of finding a spatiotemporal estimate with piecewise constant fits in the spatial dimension and piecewise constant fits in the spectral dimension is $O(N^2 M)$. To see this, first note that the computational complexity of performing binary tree pruning of a spectrum of length $M$ and fitting constants to each optimal partition interval for a single spectrum is $O(M)$ because constant fits (and their likelihoods) at one node of a tree can be computed from the average of the constant fits of the children. In the full spatiotemporal algorithm, as described in Sec. 2.2, the spectral smoothing operation is performed on every unique spectrum at each level of the quadtree. At scale $j$ of the quadtree, there are $N^2/2^{2j}$ spectra to estimate, where $j$ ranges from 0 to $\log_2 N$. Since $\sum_{j=0}^{\log_2 N} \frac{N^2}{2^{2j}} = O(N^2)$, the computational complexity of finding the optimal spatiotemporal partition and fitting piecewise constants to the partition cells is $O(N^2 M)$. 

27
The computational complexity of finding a spatiospectral estimate with piecewise constant fits in the spatial dimension and piecewise polynomial fits in the spectral dimension depends on the complexity of the optimization routine used to find the polynomial coefficients. To prune a spectrum of length $M$ and to fit polynomials to each partition interval, we will need $O(M)$ likelihood function calls (that can be computed using a total of $O(M \log_2 M)$ operations), and $O(M)$ optimization routine calls. Thus, the full spatiospectral denoising algorithm that makes piecewise constant fits in the spatial dimension and piecewise polynomial fits in the spectral dimension requires $O(N^2 M \log_2 M)$ likelihood function calls and $O(N^2 M)$ polynomial fitting routine calls.

The accuracy of the proposed estimator can be augmented by a process called cycle-spinning, or averaging over shifts, resulting in translation-invariant (TI) estimates [70, 71]. Cycle-spinning was derived in the context of undecimated wavelet coefficient thresholding in the presence of Gaussian noise, but can be difficult to implement efficiently in our case when spectral smoothing is performed at every leaf of the quadtree. If spectral smoothing is not required (which might be the case when the spectral intensity is uncorrelated along the spectral dimension), then TI estimates with piecewise constant spatial fits can be obtained in $O(N M \log_2 N)$ time using some novel computational methods discussed in [65].

2.5 Spectral image reconstruction

The proposed multiscale method for spatiospectral denoising can now be used in an Expectation-Maximization framework to reconstruct a blurred noisy spectral image. As explained in Sec. 2.1.1, the observations $\mathbf{y} \sim \text{Poisson}(\mathbf{A} \mathbf{f})$ collected at the detector are noisy and distorted, where $\mathbf{A}$ corresponds to the distortion operator, and our goal is to estimate $\mathbf{f}$ from $\mathbf{y}$ as accurately as possible.

To solve this challenging inverse problem, we perform the following optimization
problem:

\[
\hat{f} = \arg\min_{f \in \Gamma_{M,N}} \left\{ -\log p(y | A\tilde{f}) + \text{pen}(\tilde{f}) \right\},
\]

where \( \Gamma_{M,N} \) is the collection of estimators corresponding to a pruned spatio-spectral tree as described in Section 2.2, and the penalty \( \text{pen}(\tilde{f}) \) is proportional to the number of cells in the pruned RDP; hence the penalty term encourages solutions with small numbers of leaves. We compute the solution to this problem using an Expectation-Maximization algorithm, which in this case is a regularized version of the Richardson-Lucy algorithm \([14,72,73]\). The method consists of two alternating steps:

**E-step:** \( x^{(t)} = \hat{f}^{(t)} \times A^T(y./A\hat{f}^{(t)}) \), where \( \times \) and \( ./ \) denote element-wise multiplication and division, respectively.

**M-step:** Compute \( \hat{f}^{(t+1)} \) by denoising \( x^{(t)} \) as described in Section 2.2.

### 2.6 Experimental results

In this section we demonstrate the effectiveness of the proposed spatio-spectral algorithm on the NASA AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) Moffett field reflectance data set.

#### 2.6.1 Denoising results

In these experiments, we focus on a region of the data cube which is \( 256 \times 256 \times 128 \). Furthermore, we scale the data so that observations are truly photon-limited and the mean intensity per voxel is 0.0387. Figs. 2.3(a) and 2.3(b) show the true intensity and the noisy observations at spectral bands 6 respectively. Here we compare our denoising algorithm to three other approaches, a) Kolaczyk’s corrected Haar thresholds extended to spectral data \([10,11]\), b) Kolaczyk’s and Nowak’s multiscale CPLE method applied to every spectral image independently \([16]\) and c) the Fourier-
wavelet-based spectral image estimation algorithm (Atkinson method) [20]. To provide a quantitative comparison of the results, we use the following error measure:

$$\varepsilon = \frac{\|f - \hat{f}\|_1}{\|f\|_1}.$$  

In the experiments discussed below, the error numbers are obtained by averaging the error obtained by running independent experiments over 100 different noise realizations.

Kolaczyk’s corrected Haar wavelet thresholds algorithm extends the wavelet-based thresholding approaches to Poisson data. To account for the signal-dependence of the Poisson statistics, Kolaczyk suggested the use of corrected, scale-dependent thresholds at every level of decomposition based on the tail probabilities of the wavelet coefficients followed by hard or soft thresholding. Here we extend this approach to 3d Haar wavelet transform to denoise spectral data and use the following thresholds at every level $j = 0, 1, \ldots, J$ for $J = \log_2 N, \log_2 M$:

$$t_j = t'2^{-3(J-j)/2}\left[\log(n_j) + \sqrt{\log^2(n_j) + 2\lambda_j \log(n_j)}\right]$$

where $n_j = 2^{3j}$, $\lambda_j = 2^{3(J-j)}\lambda'$, and $t'$ and $\lambda'$ are user-defined parameters that are chosen to minimize the error. In this experiment, we use hard-thresholding. This approach assumes the same degrees of smoothness in the spatial and the spectral dimensions. However, if the underlying intensity has anisotropic smoothness in the spatial and the spectral dimensions, such an approach becomes less effective as demonstrated by the results shown in Figs. 2.3(c) and 2.4(a) respectively. The parameters $t'$ and $\lambda'$ are chosen to minimize the error, which is $\varepsilon = 0.3497$.

The multiscale CPLE method finds a spatial partition that maximizes the penalized log likelihood from the space of all possible recursive dyadic partitions in $[0, 1]^2$ and fits piecewise constants to every partition interval [16]. In this experiment, we weighted the spatial penalty to minimize the error. Since the observations are extremely photon-limited, performing the multiscale CPLE method on every spectral
image independently yields oversmoothed results, as shown in Figs. 2.3(d) and 2.4(b) respectively. The average error is \( \varepsilon = 0.2887 \).

![Figure 2.3: Spatiospectral denoising results. (a) True intensity at spectral band 6. (b) Noisy observations at spectral band 6. (c) Result obtained by applying the corrected 3d Haar wavelet thresholds [10, 11], seen at spectral band 6; \( \varepsilon = 0.3497 \). (d) Result obtained by performing the multiscale CPLE method [16] on every image separately, seen at spectral band 6; \( \varepsilon = 0.2887 \). (e) Result obtained by performing the Atkinson method [20], seen at spectral band 6; \( \varepsilon = 0.2298 \). (f) Result obtained by performing the proposed full spatiospectral denoising algorithm averaged over 2000 different spatial and spectral shifts, seen at spectral band 6; \( \varepsilon = 0.1917 \).](image)

The Atkinson method proposed in [20] is designed to perform spectral intensity estimation from Gaussian observations. A possible approach for handling Poisson data is to apply the Anscombe transform to the Poisson data and approximate them to Gaussian. However, such a transformation is inaccurate when the observations are extremely photon-limited; for example, when the mean intensity per voxel is 0.0387, then the number of photons measured at each voxel is either 0 or 1. Under such circumstances, the Anscombe transform breaks down and hence yields suboptimal results. In our experiments, we found that the error was much higher \( \varepsilon = 0.3471 \) when we performed the wavelet-based spectral intensity estimation algorithm after...
applying the Anscombe transform to the Poisson data. The results obtained by this algorithm without variance stabilization are shown in Figs. 2.3(e) and 2.4(c) respectively, and the error is $\varepsilon = 0.2298$. We chose a wavelet threshold that minimized the error. From the results, we can see that this algorithm performs better than the multiscale CPLE method because it accounts for the spectral correlation that exists between different spectral bands. However, the algorithm fails to preserve some fine edges when the number of observed photons is extremely low, as shown in Fig. 2.3(e).

Figs. 2.3(f) and 2.4(d) show the results obtained by applying the proposed full spatiotemporal denoising algorithm with $r = 1$, averaged over 2000 random spatial and spectral shifts to overcome the blocky artifacts introduced by the full spatiotemporal denoising algorithm ($\varepsilon = 0.1917$). The penalty in (2.4) was multiplied by a small constant factor to yield a low error and also visually appealing results. From the results, we can see that the proposed approach outperforms the other two algorithms. The algorithm is very effective in estimating fine (high-frequency) details even when the observations are extremely photon-limited, as seen in Fig. 2.3(f). Similar results were achieved using piecewise polynomial fits to the log of the spectrum using generalized linear models [64].

2.6.2 Reconstruction results

Reconstruction problems are generally much more challenging and computationally intensive than the denoising problems because of the iterative algorithms often used to solve such problems. In the experiments discussed below, we restrict our attention to a smaller subset of the AVIRIS data of size $128 \times 128 \times 64$ to reduce the computational complexity. We also increase the mean intensity per voxel to be 20.7614 because of the difficulty associated with the reconstruction in the inverse problem setting. In this experiment, we apply the full spatiotemporal denoising algorithm as the M-step of the EM algorithm and fit piecewise constants both spatially and
Figure 2.4: Spatiotemporal denoising results. True spectrum at location (31,160) is shown by a dashed black line and estimates are shown by solid gray lines. (a) The estimate obtained by using the Kolaczyk’s corrected Haar wavelet thresholding approach, (b) the estimate obtained by performing the multiscale CPLE method on every image independently, (c) the estimate obtained by using the Atkinson method, (d) the estimate obtained by using the proposed full spatiotemporal denoising algorithm.

spectrally. To overcome the blocky artifacts resulting from the full spatiotemporal denoising algorithm, and to approximate cycle-spinning, we perform the following at every iteration:

**E-step:** \( \mathbf{x}^{(t)} = \tilde{\mathbf{f}}^{(t)} \times A^T (\mathbf{y}./A\tilde{\mathbf{f}}^{(t)}) \).

**M-step:** Compute \( \tilde{\mathbf{f}}^{(t+1)} = S_i^{-1} \text{[Denoise}[S_i(\mathbf{x}^{(t)})]] \).

**Estimate at iteration** \( t \): \( \tilde{\mathbf{f}}^{(t+1)} = \frac{1}{k} \sum_{i=1}^{k} \tilde{\mathbf{f}}^{(t+1)}_i \)

The shift operator \( S_i(.) \) shifts the argument \( . \) by \( i \) voxels in either of the three coordinate directions, and the inverse operator \( S_i^{-1}(.) \) undoes the shifting. This procedure is a very good approximation to cycle-spinning as the number
of shifts \( k \) approaches the detector resolution \( N^2 M \). We stop iterating when
\[
\frac{\| \hat{f}^{(t+1)} - \hat{f}^{(t)} \|^2}{\| \hat{f}^{(t)} \|^2} < 1 \cdot 10^{-4}.
\]

Fig. 2.5(a) shows the true intensity at spectral band 61. Fig. 2.5(b) shows the blurred and noisy observations measured at the detector. Fig. 2.5(c) shows the result of the RL algorithm after convergence. The RL reconstruction is very noisy as is evident from Fig. 2.5(c); the error is \( \varepsilon = 0.2700 \). In order to provide a more meaningful comparison, we computed an ad hoc approximation to the EM algorithm, in which we used the Atkinson method of [20] in the M-step; this yielded the result presented in Fig. 2.5(d). The method is considered an ad hoc, approximate EM algorithm because it does not correspond to the penalized likelihood criterion in equation (2.18) for any known penalization method or prior probability model on
the wavelet and Fourier coefficients. In performing this experiment, we used the undecimated wavelet transform and found a threshold that minimizes the error and also yields visually sharp results. The error associated with this approach is \( \varepsilon = 0.1467 \). Fig. 2.5(e) shows the result obtained by using the proposed reconstruction algorithm. As with the previous algorithm, the spatial and the spectral penalties were chosen to yield an estimate with minimum error and visually sharp results, the error is \( \varepsilon = 0.1415 \). Visually, the reconstruction results shown in Figs. 2.5d and 2.5e are comparable. This is largely because of the fact that we use an undecimated wavelet transform in the wavelet-based spectral reconstruction algorithm and just average over nine different spatial shifts in the case of the proposed approach. To make a fair comparison, we conducted an experiment with the Atkinson method using the decimated wavelet transform and averaged over nine spatial shifts similar to our proposed approach. The reconstruction with the Atkinson method did not converge, and the results were unacceptable.

In all the experiments discussed above, the penalties were chosen to minimize the error and yield visually appealing results since the ground truth was available. When the true spectral data are unknown, but multiple realizations of the data are available, then the data can be split to test and validation data, and methods such as cross validation can be used to find the tuning parameters.

In related work, our approach was shown to be similarly effective in the context of compressed sensing of spectral images. Many modern spectral imagers face a limiting tradeoff between spatial and spectral resolution, with the total number of voxels measured constrained by the size of the detector array. To mitigate this tradeoff, many researchers have developed spectral imaging systems and associated reconstruction methods that are designed to exploit the theory of compressive sensing [30, 31, 33, 46]. One example physical system collects coded projections of each spectrum in the spectral image [31]. Using the novel multiscale representation of
the spectral image based upon adaptive partitions as described in this work, we are able to accurately reconstruct spectral images with an order of magnitude more reconstructed voxels than measurements.

2.7 Application of our method on real, photon-limited astronomical data

Our spatiotemporal denoising algorithm has been successfully used for denoising astronomical data of galactic supernova remnant G1.9+0.3, collected by the Chandra X-Ray observatory [74, 75]. The supernova data consists of 256 spectral channels and 256 × 256 spatial locations per spectral channel. This data is extremely photon-limited; the mean intensity per voxel is 0.002994. Fig. 2.6(a) shows the noisy observations summed across the spectral channels and Fig. 2.6(b) shows the estimate obtained using our algorithm. Since the ground truth in this experiment is unknown, the optimal spatial and spectral penalties are found by performing cross validation. These results are used to study the expansion of galactic supernova remnant G1.9+0.3 over a period of few years [74].

2.8 Discussion

In this chapter, we presented an efficient multiscale algorithm for estimating spectral Poisson intensities. The key feature of our algorithm is that it adapts to varying degrees of smoothness in the spatial and the spectral directions unlike 3d wavelet transform based approaches that assume the same degree of smoothness in all coordinate directions. Furthermore, our approach is backed by strong theoretical results which show that the proposed spatiotemporal algorithm is nearly minimax optimal on a certain anisotropic Hölder-Besov function class. The experimental results suggest that the algorithm performs very well even when there are significantly fewer photons than voxels. In many practical applications, the observed data are binned because
Figure 2.6: Denoising results on real Chandra data. (a) Photon-limited data summed across the spectral channels. Mean photon intensity per voxel is 0.002994. (b) Estimate obtained using our algorithm where the optimal spatial and spectral penalty values are obtained by cross validation. The estimate is color coded to indicate different spectral regions; Red, 1 to 3 keV; green, 3 to 4.5 keV; blue, 4.5 to 7.5 keV. This figure is taken from [74].

of the inability of the reconstruction software to reconstruct the true intensity from such low photon counts. In contrast, our algorithm can offer improved estimation accuracy even when the observations are photon-limited. The near-optimality of our algorithm on a wide range of spectral intensities and the superior performance in photon-limited scenarios suggest that it can be an important component of several applications.
This chapter describes computationally efficient approaches and associated theoretical performance guarantees for the detection of known spectral targets and spectral anomalies from few projection measurements of spectral images. The proposed approaches accommodate spectra of different signal strengths contaminated by a colored Gaussian background, and perform detection without reconstructing the spectral image from the observations. The theoretical performance bounds of the target detector highlight fundamental tradeoffs among the number of measurements collected, amount of background signal present, signal-to-noise ratio, and similarity among potential targets in a known spectral dictionary. The anomaly detector is designed to control the number of false discoveries below a desired level and can be adapted to uncertainties in the user's knowledge of the spectral dictionary. Unlike approaches based on the principles of compressed sensing, the proposed approach does not depend on a known sparse representation of targets; rather, the theoretical performance bounds exploit the structure of a known dictionary of targets and the
distance preservation property of the measurement matrix. Simulation experiments illustrate the practicality and effectiveness of the proposed approaches.

3.1 Spectral target and anomaly detection

Spectral imaging is exploited in several applications since it yields spectral information for every spatial location in a scene that enables one to characterize the material composition of the underlying environment. In remote sensing, such images are often used in two different contexts: (1) to study a scene of interest by classifying every object in the scene to different known classes [7], and (2) to detect man-made targets against a natural landscape for surveillance applications [8]. In ophthalmology, hemoglobin spectral signatures provide information about the oxygen supply to the eye, which in turn helps to identify any tissue damage [76].

Accurately detecting target or anomalous spectra is complicated by a variety of factors:

- The resolution of conventional spectral images is typically limited because of physical and cost considerations.
- Each measured spectrum reflects the mixing of multiple spectra across an area proportional to the system resolution.
- Dictionary of spectra collected in a laboratory does not always reflect real-world operating conditions.

To see the impact of resolution limits, consider the following toy example. Let $f_1$ and $f_2$ be two spectra (functions of wavelength $\lambda$) defined as follows:

$$f_1(\lambda) = \alpha \left( \mathbb{I}_{\lambda \in [0,s/2)} - \mathbb{I}_{\lambda \in [s/2,s)} \right), f_2(\lambda) = 0$$

where $\alpha > 0$ is an amplitude factor, $\mathbb{I}_A = \{1 \text{ if } A \text{ is true and } 0 \text{ otherwise}\}$ is an indicator function, and $s \in [0,1]$. Note that $f_1$ and $f_2$ are identical except over an
interval of length \( s \), and the \( L_2 \) distance between them is \( \alpha \sqrt{s} \). Now consider collecting low-resolution observations of \( f_1 \) and \( f_2 \) by integrating the observations over \( K \) uniform intervals of width \( 1/K > s \). The distance between these low-resolution spectra will be zero. That is, these spectra will be indistinguishable unless the resolution of the detector (i.e., the number of samples \( K \)) is high with respect to \( s \). Thus target detection with conventional spectral imagers is limited by the resolution of the detectors. In this work, we present an alternative measurement paradigm based on projective measurements (discussed in detail in Sec. 3.2) which would be capable of distinguishing \( f_1 \) and \( f_2 \) even for relatively small \( K \). In contrast, a single random projection of each of these spectra would suffice to show that they are different. This leads to the conclusion that spectral target detection may benefit greatly from such projection measurements when the number of measurements that can be collected \( (K) \) is small.

Recent advances in compressive sensing (CS) have shown that it is possible to accurately reconstruct a sparse signal from a few (relative to the signal dimension) projections \([28, 29]\). This work has led to the development of new spectral imaging platforms which attempt to address challenges related to system size, resolution, and noise by acquiring fewer compressive measurements than spatiotemporal voxels \([30–35]\). However, these system designs have a number of degrees of freedom which influence subsequent data analysis. For instance, the single-shot compressive spectral imager discussed in \([31]\) collects one coded projection of each spectrum in the scene. One projection per spectrum is sufficient for reconstructing spatially homogeneous spectral images, since projections of neighboring locations can be combined to infer each spectrum. Significantly more projections are required for detecting targets of unknown strengths without the benefit of spatial homogeneity. We are interested in investigating how several such systems can be used in parallel to reliably detect spectral targets and anomalies from different coded projections.
In this work we develop detection performance bounds which show how performance scales with the number of detectors in a compressive setting as a function of SNR, the similarity between potential targets in a known dictionary, and their prior probabilities. Our bounds are based on a detection strategy which operates directly on the collected data as opposed to first reconstructing a spectral image and then performing detection on the estimated spectral image. Reconstruction as an intermediate step in detection may be appealing to end users who wish to visually inspect spectral images instead of relying entirely on an automatic detection algorithm. However, using this intermediate step has two potential pitfalls. First, the Rao–Blackwell theorem [77] tells us that an optimal detection algorithm operating on the processed data (i.e. not sufficient statistics) cannot perform better than an optimal detection algorithm operating on the raw data. In other words, optimal performance is possible on the raw data, but we have no such performance guarantee for the reconstructed spectral image. Second, the relationship between reconstruction errors and detection performance is not well understood in many settings. Although we do not reconstruct the full spectral image, our performance bounds are intimately related to the spectral image resolution needed to achieve the spectral diversity present in our dictionary. Since we have many fewer observations than the spectral image at this resolution, we adopt the “compressive” terminology.

Parts of this chapter have been adapted from [78] and [79] with permissions from IEEE.

3.1.1 Problem formulation

Let us assume access to a spectral dictionary \( \mathcal{D} = \{\mathbf{f}^{(1)}, \mathbf{f}^{(2)}, \ldots, \mathbf{f}^{(m)}\} \), where \( \mathbf{f}^{(j)} \in \mathbb{R}^N \) for \( j = 1, \ldots, m \) is unit-norm. Our measurements are of the form

\[
z_i = \Phi(\alpha_i \mathbf{f}_i^* + \mathbf{b}_i) + \mathbf{w}_i \tag{3.1}
\]

where
• $i \in \{1, \ldots, M\}$ indexes the spatial locations of the spectral data;

• $\alpha_i \geq 0$ is a measure of the signal-to-noise ratio at location $i$, which is either known or estimated from observations;

• $\Phi \in \mathbb{R}^{K \times N}$ for $K < N$, is a measurement matrix to be specified in Sec. 3.2;

• $b_i \in \mathbb{R}^N \sim \mathcal{N}(\mu_b, \Sigma_b)$ is the background noise vector;

• $w_i \in \mathbb{R}^K \sim \mathcal{N}(0, \sigma^2 I)$ is the i.i.d. sensor noise.

In this work we consider the following spectral target detection problems:

1. Dictionary spectrum detection (DSD): Here we assume that each observed spectrum comes from a known spectral dictionary, and our task is to detect all instances of one dictionary element across the spectral image, i.e. $f_i^* = f^{(j)} \in \mathcal{D}$ for some (unknown) $j \in \{1, \ldots, m\}$. DSD is useful in contexts in which we know the spectral makeup of a scene and wish to focus our attention on the locations of a particular spectrum [80].

2. Anomalous spectrum detection (ASD): Here, our task is to detect all spectra which are not members of our spectral dictionary, i.e. $f_i^* \notin \mathcal{D}$. (This is akin to anomaly detection methods in the literature which are based on nominal, non-anomalous training samples [81, 82].) ASD is useful when we know most spectral components of a scene and wish to identify all spectra which deviate from this model [83].

Our goal is to accurately perform DSD or ASD without reconstructing the spectral input $f_i^*$ from $z_i$ for $i \in \{1, \ldots, M\}$. Typically, in spectral target detection, the background corresponding to the scene of interest and the sensor noise are modeled together by a colored multivariate Gaussian distribution [84]. However, in our case,
it is important to distinguish the two because of the presence of the projection operator $\Phi$. The projection operator acts upon the background spectrum in the same way as on the target spectrum, but it does not affect the sensor noise. We assume that $b_i$ and $w_i$ are independent of each other, and the prior probabilities of different targets in the spectral dictionary $p^{(j)} = P(f_i^* = f^{(j)})$ for $j \in \{1, \cdots, m\}$ are known in advance. If these probabilities are unknown, then the targets can be considered equally likely. Given this setup, our goal is to develop suitable target and anomaly detection approaches, and provide theoretical guarantees on their performance.

3.1.2 Performance metric

To assess the performance of our detection strategies, we consider the False Discovery Rate (FDR) metric and related quantities developed for multiple hypothesis testing problems [37]. Since our spectral image has $M$ spatial locations, we are simultaneously conducting $M$ hypothesis tests when we search for targets. Unlike the probability of false alarm, which measures the probability of falsely declaring a target at a single location, the FDR measures the fraction of declared targets that are false alarms, that is, it provides information about the entire set of $M$ hypotheses instead of just one. More formally, the FDR is given by

$$
FDR = \mathbb{E} \left[ \frac{V}{R} \right],
$$

where $V$ is the number of falsely rejected null hypotheses, and $R$ is the total number of rejected null hypotheses. Controlling the false discovery rate in a multiple hypothesis testing framework is akin to designing a constant false alarm rate (CFAR) detector in spectral target detection applications that keeps the false alarm rate at a desired level irrespective of the background interference and sensor noise statistics [80]. Although FDR has not been used extensively in the spectral imaging community, it has been widely adopted in other research communities where multiple hypothesis testing is
common, including biostatistics and fMRI activity detection [37, 85–87].

3.1.3 Previous investigations

Target detection in spectral images

There is a rich literature on the detection of spectral targets and anomalies from spectral data [80, 83, 84, 88–98]. Usually, the targets are assumed to occupy a small fraction of the entire image scene. The differences in the target detection approaches among competing methods arise from various models of the target and the background statistics, and the composition of each pixel in the scene. The full pixel model assumes that each spatial location contains either a target or background material. Under this model, target detection is performed using likelihood ratio tests (LRT) if the target and background statistics are known accurately, or using generalized likelihood ratio tests (GLRT) if they are estimated from the data [80, 84]. Other detection methods for the full pixel model include the spectral angle mapper (SAM) which detects spectral targets by measuring the similarity between the observed spectrum and the known target spectrum using the cosine of the angle between the two spectra [88].

The full pixel model is reasonable if the spatial resolution is high enough to guarantee that there is only one object present at any given pixel. However, in many applications, the spatial area covered by each pixel might contain multiple objects, and have background interference from the surrounding materials. A mixed pixel or subpixel model accounts for such interferences by modeling every spatial location as either a target material corrupted by background, or just background [80, 84]. In our problem formulation, we use a multivariate normal distribution to account for the spectral variability of the possible targets of interest and the background interference due to the aggregation of different objects at every spatial location (e.g. [80]).

In a mixed pixel model, the target spectra are often assumed to lie in a $P$-
dimensional subspace of $\mathbb{R}^N$ for $P < N$ (subspace model), where the dimension of the target subspace grows with the variability of the target spectra [90–92,95,96]. A linear mixing model is a special case of a subspace model where the target subspace is spanned by spectra from a known spectral library whose relative abundances are restricted to be nonnegative and sum to one [80,84]. The subspace in which the target lies is assumed known or specified by the user in a subspace model, and the variability of the background is modeled using a probability distribution. Given knowledge of the target subspace, background statistics and sensor noise statistics, detection methods based on LRTs and GLRTs have been proposed [89–92,95,96]. Recent studies have shown that spectral target detection and spectral demixing can be efficiently performed by solving an $\ell_1$ minimization problem that finds spectra in the observed data that match the target spectrum of interest subject to the constraint that the number of pixels in the data that contain the target spectrum is small [93,94].

Similar approaches based on GLRT have also been proposed in anomaly detection problems where the target spectra of interest are not known a priori [83,92,97,98]. Here the target spectra are either assumed to be drawn from a multivariate normal distribution whose parameters are unknown [97], or considered to lie in a known subspace [92]. The variability of the background is modeled in a similar fashion, either using a subspace model or a statistical model. Since it is impossible to estimate the covariance matrices of the anomalous targets, it is assumed that the anomalous spectra and the background have the same covariances. The GLRT-based anomaly detection algorithms typically model the background as globally or locally homogeneous and identify the spatial locations that contain spectral signatures that are significantly different from the known or estimated background [83]. While these approaches are shown to work well in practice [80,83,84], extending them to more general sampling schemes is not straightforward.
Detection and classification from projection measurements

There exist several methods for target or anomaly detection that rely on recovering the full spatiotemporal data from projection measurements [45, 99]. However, they are computationally intensive and the detection performance associated with these reconstructions is unknown. Other researchers have exploited compressive sensing to perform target detection and classification without reconstructing the underlying signal [100–102]. In [102], the authors propose a matching pursuit based algorithm, called the *Incoherent Detection and Estimation Algorithm* (IDEA), to detect the presence of a signal of interest against a strong interfering signal from noisy projection measurements. The algorithm is shown to perform well on experimental data sets under some strong assumptions on the sparsity of the signal of interest and the interfering signal. However, there are no theoretical guarantees on the performance of this detector relative to the number of measurements that one collects, and it is not clear how this approach can be extended to anomaly detection problems. In [100], the authors develop a classification algorithm called the *smashed filter* to classify an image in \( \mathbb{R}^N \) to one of \( m \) known classes from \( K \) projections of the signal, where \( K < N \). The underlying image is assumed to lie on a low-dimensional manifold, and the algorithm finds the closest match from the \( m \) known classes by performing a nearest neighbor search over the \( m \) different manifolds. The projection measurements are chosen to preserve the distances among the manifolds. Though [100] offers theoretical bounds on the number of measurements necessary to preserve distances among different manifolds, it is not clear how the performance scales with \( K \) or how to incorporate background models into this setup. Moreover, this approach may be computationally intensive since it involves learning and searching over different manifolds. In [101], the authors use a nearest-neighbor classifier to classify an \( N \)-dimensional signal to one of \( m \) equally likely target classes based
on $K < N$ random projections, and provides theoretical guarantees on the detector performance. While the method discussed in [101] is computationally efficient, it is nontrivial to extend to the case of spectral target detection with colored background noise and non-equiprobable targets. Furthermore, their performance guarantees cannot be directly extended to our problem since we focus on error measures that let us analyze the performance of multiple hypothesis tests simultaneously as opposed to the above method that considers compressive classification performance at a single spatial location.

3.1.4 Contributions

This work makes the following contributions to the above literature:

- A compressive spectral target detection approach, which (a) is computationally efficient, (b) allows for the signal strengths of the targets to vary with spatial location, (c) allows for spectral backgrounds mixed with potential targets, (d) considers targets with different a priori probabilities, and (e) yields theoretical guarantees on detector performance.

- A computationally efficient anomaly detection method that detects spectral anomalies of different strengths from projection measurements and also controls the false discovery rate at a desired level.

- A whitening filter approach to compressive measurements of signals with background contamination, and associated analysis leading to bounds on the amount of background to which our detection procedure is robust.

The above theoretical results, which are the main focus of this work, are supported with simulation studies in Sec. 3.6. Classical detection methods described in [80, 83, 84, 88–98] do not establish performance bounds as a function of spectral resolution or dictionary properties and rely on relatively direct observation models which we
show to be suboptimal when the detector size is limited. The methods in [100]
and [102] do not contain performance analysis, and our analysis builds upon the
analysis in [101] to account for several specific aspects of the spectral image target
detection problem.

3.2 Whitening compressive observations

Before we present our detection methods for DSD and ASD problems respectively,
we briefly discuss a whitening step that is common to both our problems of interest.

Let us suppose that there are enough background training data available to esti-
mate the background mean \( \mu_b \) and covariance matrix \( \Sigma_b \). We can assume without
loss of generality that \( \mu_b = 0 \) since \( \Phi \mu_b \) can be subtracted from \( y \). Given the knowl-
dge of the background statistics, we can transform the background and sensor noise
model \( \Phi b_i + w_i \sim \mathcal{N}(0, \Phi \Sigma_b \Phi^T + \sigma^2 I) \) discussed in (3.1) to a simple white Gaussian
noise model by multiplying the observations \( z_i, i \in \{1, \ldots, M\} \), by the whitening
filter

\[
C_\Phi \triangleq (\Phi \Sigma_b \Phi^T + \sigma^2 I)^{-1/2}.
\]

This whitening transformation reduces the observation model in (3.1) to

\[
y_i = C_\Phi (\Phi (\alpha_i f_i^* + b_i) + w_i) = \alpha_i A f_i^* + n_i \tag{3.2}
\]

where

\[
A = C_\Phi \Phi,
\]

and \( n_i = C_\Phi (\Phi b_i + w_i) \sim \mathcal{N}(0, I) \). To verify that \( n_i \sim \mathcal{N}(0, I) \), observe that

\[
n_i = C_\Phi (\Phi b_i + w_i) \sim \mathcal{N}(0, C_\Phi (\Phi \Sigma_b \Phi^T + \sigma^2 I) C_\Phi^T).
\]

We can now choose \( \Phi \) so that the corresponding \( A \) has certain desirable properties
as detailed in Sec. 3.3 and Sec. 3.5.
For a given $A$, the following theorem provides a construction of $\Phi$ that satisfies (3.3) and a bound on the maximum tolerable background contamination:

**Theorem 3.** Let $B = I - A\Sigma_b A^T$. If the largest eigenvalue of $\Sigma_b$ satisfies

$$\lambda_{\text{max}} < \frac{1}{\|A\|^2},$$

(3.4)

where $\|A\|$ is the spectral norm of $A$, then $B$ is positive definite and $\Phi = \sigma B^{-1/2} A$ is a sensing matrix, which can be used in conjunction with a whitening filter to produce observations modeled in (3.2).

The proof of this theorem is provided in Appendix A.2.1. This theorem draws an interesting relationship between the maximum background perturbation that the system can tolerate and the spectral norm of the measurement matrix, which in turn varies with $K$ and $N$. Hardware designs such as that in [33, 35] use spatial light modulators and digital micro mirrors, which allow the measurement matrix $\Phi$ to be adjusted easily in response to changing background statistics and other operating conditions.

In the sections that follow, we consider collecting measurements of the form $y_i = \alpha_i A f_i^* + n_i$ given in (3.2), where $f_i^*$ is the target of interest for $i = 1, \ldots, M$, and $A \in \mathbb{R}^{K \times N}$ is a sensing matrix that satisfies (3.3). It is assumed that any background contamination has been eliminated with the whitening procedure described in this section.

### 3.3 Dictionary spectrum detection

Suppose that the end user wants to test for the presence of one known target versus the rest, but it is not known a priori which target from $D$ the user wants to detect. In this case, let us cast the DSD problem as a multiple hypothesis testing problem...
of the form
\[
\mathcal{H}^{(j)}_{0i} : f_i^* = f^{(j)} \quad \text{vs.} \quad \mathcal{H}^{(j)}_{1i} : f_i^* \neq f^{(j)}
\]  \hspace{1cm} (3.5)

where \( f^{(j)} \in D \) is the target of interest and \( i = 1, \ldots, M \).

3.3.1 Decision rule

We define our decision rule corresponding to target \( f^{(j)} \in D \) in terms of a set of significance regions \( \Gamma_i^{(j)} \) such that one rejects the \( i \)-th null hypothesis if its test statistic \( y_i \) falls in the \( i \)-th significance region. Specifically, \( \Gamma_i^{(j)} \) is defined according to

\[
\Gamma_i^{(j)} = \left\{ y : \log P(f_i^* = f^{(j)} | y, \alpha_i, A) \leq \right. \\
\log P(f_i^* = f^{(\ell)} | y, \alpha_i, A) \text{ for some } \ell \in \{1, \ldots, m\}, \ell \neq j \},
\]

where

\[
\log P(f_i^* = f^{(j)} | y, \alpha_i, A) = \frac{K}{2} \log \left( \frac{1}{2\pi} \right) - \frac{\|y - \alpha_i Af^{(j)}\|^2}{2} + \log p^{(j)}
\]

is the logarithm of the a posteriori probability density of the target \( f^{(j)} \) at the \( i \)-th spatial location given the observations \( y \), the signal-to-noise ratio \( \alpha_i \) and the sensing matrix \( A \), and \( p^{(j)} \) is the a priori probability of target class \( j \). The decision rule can be formally expressed in terms of the significance regions as follows:

\[
\text{reject } \mathcal{H}^{(j)}_{0i} \text{ if the test statistic } y_i \in \Gamma_i^{(j)}.
\]  \hspace{1cm} (3.7)

We analyze this detector by extending the positive False Discovery Rate (pFDR) error measure introduced by Storey to characterize the errors encountered in performing multiple, independent and non-identical hypothesis tests simultaneously [36]. The pFDR, discussed formally below, is the fraction of falsely rejected null hypotheses among the total number of rejected null hypotheses, subject to the positivity
condition that one rejects at least one null hypothesis. The pFDR is similar to the FDR except that the positivity condition is enforced here. In our context, the positivity condition means that we declare at least one spectrum to be a nontarget, which in turn implies that the scene of interest is composed of more than one object.

Suppose that we perform $M$ identical and independent hypothesis tests simultaneously, and reject the $i^{th}$ null hypothesis if the test statistic $y_i$ corresponding to the $i^{th}$ hypothesis test falls in a fixed, predetermined significance region $\Gamma_i$. For identical hypothesis tests, the significance regions corresponding to each test will be the same, and the pFDR in this case is given by

$$pFDR(\Gamma) = E \left[ \frac{V(\Gamma)}{R(\Gamma)} \middle| R(\Gamma) > 0 \right]$$

(3.8)

where $V(\Gamma)$ is the number of falsely rejected null hypotheses, and $R(\Gamma)$ is the total number of rejected null hypotheses. Storey proved that if the tests are identical and independent, the pFDR can simply be written in terms of posterior probability as

$$pFDR(\Gamma) = P(\mathcal{H}_0 | y \in \Gamma)$$

(3.9)

where the subscript $i$ is dropped to signify that $P(\mathcal{H}_{0i} | y_i \in \Gamma_i)$ is the same for $i = 1, \ldots, M$ because the tests are (assumed to be) identical and independent [36]. This pFDR expression is particularly useful since the probability term in (3.9) can often be upper-bounded analytically.

In our setup, however, the hypothesis tests are independent but nonidentical, since the significance regions in (3.6) depend on the signal-to-noise ratio $\alpha_i$ at every spatial location. Nevertheless, the pFDR error measure in (3.8) can be suitably modified and expressed in terms of posterior probability as given in (3.9), to accommodate the testing of such multiple, nonidentical hypothesis tests simultaneously.

Consider a collection of significance regions $\Gamma = \{\Gamma_i^{(j)} : i = 1, \ldots, M\}$, such that one declares $\mathcal{H}_{0i}^{(j)}$ if the test statistic $y_i \in \Gamma_i^{(j)}$. The pFDR for multiple, nonidentical
hypothesis tests can be defined in terms of the significance regions as follows:

\[
pFDR^{(j)}(\Gamma) = E \left[ \frac{V(\Gamma)}{R(\Gamma)} \right] \quad \text{subject to} \quad R(\Gamma) > 0 \]

(3.10)

where

\[
V(\Gamma) = \sum_{i=1}^{M} \mathbb{1}\{y_i \in \Gamma^{(j)}_i\} \mathbb{1}\{H_0\}
\]

(3.11)

is the number of falsely rejected null hypotheses,

\[
R(\Gamma) = \sum_{i=1}^{M} \mathbb{1}\{y_i \in \Gamma^{(j)}_i\}
\]

(3.12)

is the total number of rejected null hypotheses, and \(\mathbb{1}(E) = 1\) if event \(E\) is true and 0 otherwise. In our setup, the pFDR corresponds to the expected ratio of the number of missed targets to the number of spectra declared to be nontargets subject to the condition that at least one spectrum is declared to be a nontarget. (Note that this ratio is traditionally referred to as the positive false nondiscovery rate (pFNR), but is technically the pFDR in this context because of our definitions of the null and alternate hypotheses.)

Theorem 4. Given observations of the form in (3.2), if one performs multiple, independent, nonidentical hypothesis tests of the form (3.5) and decides according to (3.7), then the worst-case pFDR given by

\[
pFDR_{\text{max}} = \max_{j \in \{1, \ldots, m\}} pFDR^{(j)}(\Gamma),
\]

satisfies the following bound:

\[
pFDR_{\text{max}} \leq \max \left( 0, \frac{(P_e)_{\text{max}}}{1 - p_{\text{max}} - (P_e)_{\text{max}}} \right)
\]

(3.13)

where

\[
p_{\text{max}} = \max_{j \in \{1, \ldots, m\}} p^{(j)};
\]

\[
(P_e)_{\text{max}} = \max_{i \in \{1, \ldots, M\}} P(\hat{f}_i \neq f_i^*) \quad \text{and}
\]

\[
\hat{f}_i = \arg \max_{f \in \mathcal{D}} P(f_i^* = f | y_i, \alpha_i, A).
\]

(3.14)
The proof of this theorem is detailed in Appendix A.2.2. A key element of our proof is the adaptation of the techniques from [36] to nonidentical independent hypothesis tests. This theorem shows that, for a given \( A \), the worst-case pFDR is bounded from above by a function of the worst-case misclassification probability, which will generally depend of the number of observations \( K \), signal strengths \( \{\alpha_i\}_{i=1}^{M} \), similarity among different targets of interest, and a priori target probabilities.

### 3.3.2 Special case of Gaussian random matrix

In this section we derive the worst-case pFDR upper bound for the special case of a Gaussian random matrix where the entries of \( A \) are i.i.d. draws from \( \mathcal{N}(0, 1/K) \), and discuss how the bound behaves as a function of \( K \), \( \{\alpha_i\}_{i=1}^{M} \), and the a priori target probabilities. Specifically, we derive (a) a sufficient condition for the bound of Theorem 3 to hold with high probability in terms of the dimensions of \( A \), and (b) an upper bound on the worst-case misclassification probability and consequently, the pFDR.

**Corollary 5.** When the entries of \( A \) are drawn from \( \mathcal{N}(0, 1/K) \), there exists an absolute constant \( c > 0 \) such that the conditions of Theorem 3 are met with high probability if

\[
\lambda_{\text{max}} < \frac{1}{c\left(\sqrt{\frac{N}{K}} + 1\right)^2}.
\]

The proof of this corollary is given in Appendix A.2.3. For a given \( N \), the upper bound on \( \lambda_{\text{max}} \) increases as \( K \) increases, which implies that the system can tolerate more background perturbation if we collect more measurements.

Observe from Theorem 4 that the worst-case pFDR is upper-bounded by a function of the worst-case misclassification probability. For this particular choice of \( A \),
the corollary below finds an upper bound on the worst-case pFDR by explicitly bounding the worst-case misclassification probability as a function of $K$, $d_{\min}$ and $\{\alpha_i\}$:

**Corollary 6.** When the entries of $A$ are drawn from $\mathcal{N}(0, 1/K)$, the worst-case pFDR bound in (3.13) reduces to

$$p_{\text{FDR}}_{\max} \leq \frac{1}{p_{\min}} \left( \frac{1 - p_{\max}}{1 - p_{\min}} \left( 1 + \frac{\alpha_{\min}^2 d_{\min}^2}{4K} \right) \frac{K^2}{\pi} - \frac{1}{p_{\min}} \right)^{-1}. \quad (3.16)$$

This bound is nonnegative and less than one as long as

$$K > \frac{2 \log \left( \frac{2}{p_{\min} - p_{\max}} \right)}{\log \left( 1 + \frac{\alpha_{\min}^2 d_{\min}^2}{4K} \right)}. \quad (3.17)$$

The proof of this corollary is given in Appendix A.2.4. This corollary relates the performance of the classifier to the number of measurements $K$, the signal strength of the target at the $i$th spatial location, the similarity among different targets in $D$, and the minimum value of the a priori target probabilities. Note that the bound decays with the increase in the values of $K$, $d_{\min}$ and $\alpha_i$ and increases as $p_{\min}$ decreases. For a fixed $p_{\max}$, $p_{\min}$, $\alpha_{\min}$ and $d_{\min}$, the bound in (3.16) enables one to choose a value of $K$ to guarantee a desired pFDR value.

Observe that this bound is independent of $N$, and is only a function of $K$, $p_{\max}$, $p_{\min}$, $\alpha_{\min}$, and $d_{\min}$. The lack of dependence on $N$ is not unexpected. Indeed, when we are interested in preserving pairwise distances among the members of a fixed dictionary of size $m$, the Johnson–Lindenstrauss lemma [103] says that, with high probability, $K = \mathcal{O}(\log m)$ random Gaussian projections suffice, regardless of the ambient dimension $N$. This is precisely the regime we are working with here. The bound on $K$ given in (3.17) increases logarithmically with the increase in the difference between $p_{\max}$ and $p_{\min}$. This is to be expected since one would need more
measurements to detect a less probable target as our decision rule weights each target by its a priori probability. If all targets are equally likely, then $p_{\text{max}} = p_{\text{min}} = 1/m$, and $K = O(\log m)$ is sufficient provided $\alpha^2_{\text{min}} d^2_{\text{min}}$ is sufficiently large such that

$$\log \left( 1 + \frac{\alpha^2_{\text{min}} d^2_{\text{min}}}{4K} \right) > \log \left( 1 + \frac{\alpha^2_{\text{min}} d^2_{\text{min}}}{4N} \right) > 1$$

(where the first inequality holds since $K < N$). In addition, the lower bound on $K$ also illustrates the interplay between the signal strength of the targets, the similarity among different targets in $D$, and the number of measurements collected. A small value of $d_{\text{min}}$ suggests that the targets in $D$ are very similar to each other, and thus $\alpha_{\text{min}}$ and $K$ need to be high enough so that similar spectral targets can still be distinguished. The experimental results discussed in Sec. 3.6 illustrate the tightness of the theoretical results discussed here.

### 3.4 Extension to a manifold-based target detection framework

The DSD problem formulation in Sec. 3.1.1 is accurate if the spectra in the dictionary are faithful representations of the target spectra that we observe. In reality, the observed spectrum of any material will not match the reference spectrum of the same material observed in a laboratory because of the differences in atmospheric and illumination conditions. To overcome this problem, one could form a large spectral dictionary to account for such uncertainties in the target spectra and perform target detection according to the approaches discussed in Sec. 3.2 and Sec. 3.3. A potential drawback with this approach is that our theoretical performance bound increases with the size of $D$ through $p_{\text{min}}$ and $d_{\text{min}}$. It has been shown [104] that the set of spectra in $\mathbb{R}^N$ corresponding to a target material observed under different illumination and atmospheric conditions is a low-dimensional submanifold of the $N$-dimensional ambient signal space. We can exploit this result to extend our analysis to a much broader framework that accounts for uncertainties in our spectral dictionary.
Let us consider a dictionary of manifolds \( D_M = \{ \mathcal{M}^{(1)}, \ldots, \mathcal{M}^{(m)} \} \) corresponding to \( m \) different target classes, and that the target spectrum \( f^* \in \mathcal{M}^* \in D_M \).

Considering an observation model of the form given in (3.2), our goal is to determine if \( f_i^* \in \mathcal{M}^{(j)} \) for \( i = 1, \ldots, M \), where \( j \) is the target class of interest. Let us assume that all target classes are equally likely to keep the presentation simple, though the analysis extends to the case where the targets classes have different a priori probabilities. A related work [100] considers this manifold search problem from compressive measurements, and proposes a two-step approach to perform manifold-based classification:

1. In each manifold, find the closest point to the observed data:

   \[
   \tilde{f}_i^{(\ell)} = \arg\max_{f \in \mathcal{M}^{(\ell)}} \mathbb{P}(y_i \mid f^* = f, \alpha_i, A)
   \]

   for \( \ell \in \{1, \ldots, m\} \) and \( i = 1, \ldots, M \).

2. Form a data-dependent spectral dictionary \( D_{y_i} = \{ \tilde{f}_i^{(1)}, \ldots, \tilde{f}_i^{(m)} \} \) for every observed spectrum. Let

   \[
   \hat{f}_i = \arg\max_{f \in D_{y_i}} \mathbb{P}(y_i \mid f^* = \tilde{f}, \alpha_i, A).
   \]

   The \( i \)th observed spectrum corresponds to class \( j \) if \( \hat{f}_i = \tilde{f}_i^{(j)} \).

The second step in the two-step approach is very similar to our DSD target detection approach except that we consider a data-independent spectral dictionary in our analysis. Nevertheless, our analysis and the theoretical performance bounds extend directly to the manifold-based target detection if we collect two sets of observations \( y^{(1)} \) and \( y^{(2)} \), such that one forms data-dependent spectral dictionaries \( \{ D_{y_i^{(1)}} \} \) according to Step 1 using \( y^{(1)} \) and performs target detection on \( y^{(2)} \) using the spectral
dictionaries $\{D_y^{(1)}\}$. The hypothesis tests corresponding to the second step can be written as

$$H_0: f_i^* = \tilde{f}_i^{(j)} \text{ vs. } H_1: f_i^* \neq \tilde{f}_i^{(j)}$$

where $\tilde{f}_i^{(j)} \in D_y^{(1)}$ for $i = 1, \ldots, M$. Since the spectral dictionary in this case changes with $i$, these tests are nonidentical. This is another instance where our extension of pFDR-based analysis towards simultaneous testing of multiple, independent, and nonidentical hypothesis tests (3.10) is very significant. Following the proof techniques discussed in the appendix, we can straightforwardly show that the bound in (3.16) in this manifold setting holds with $p_{\min} = p_{\max} = 1/m$ since all target classes are assumed to be equally likely here, and $d_{\min} = \min_{i \in \{1, \ldots, M\}} d_i$ where

$$d_i = \min_{f_i^{(\ell)}, f_i^{(k)} \in D_y^{(1)}, \ell \neq k} \| \tilde{f}_i^{(\ell)} - \tilde{f}_i^{(k)} \|.$$

3.5 Anomalous spectrum detection

The target detection approach discussed above assumes that the target spectrum of interest resides in a spectral dictionary that is available to the user. However, in some applications (such as military applications and surveillance), one might be interested in detecting man-made targets against a natural landscape, whose spectral signatures are significantly different from the known spectra in the available dictionary. Since such targets often occupy a very small portion of the scene, it is difficult to collect enough training samples of every target spectrum of interest to include in our spectral dictionary. In other words, the target spectra of interest are anomalous and are not available to the user. In this section we show how the target detection methods discussed above can be extended to anomaly detection. In particular, we exploit the distance preservation property of the sensing matrix $A$ to detect anomalous targets from projection measurements.
3.5.1 Problem formulation

Given observations of the form in (3.2), we are interested in detecting whether \( f^* \in \mathcal{D} \) or \( f^* \) is anomalous. Let us write the anomaly detection problem as the following multiple hypothesis test:

\[
\begin{align*}
\mathcal{H}_{0i} : \| f^*_i - f \| &\leq \tau \text{ for some } f \in \mathcal{D} & (3.18a) \\
\mathcal{H}_{1i} : \| f^*_i - f \| &> \tau \text{ for all } f \in \mathcal{D} & (3.18b)
\end{align*}
\]

where \( \tau \in [0, \sqrt{2}) \) is a user-defined threshold that encapsulates our uncertainty about the accuracy with which we know the spectral dictionary.\(^1\) In particular, \( \tau \) controls how different a spectrum needs to be from every dictionary element to truly be considered anomalous. In the absence of any prior knowledge on the targets of interest, \( \tau \) can simply be set to zero.

Note that the definition of the hypotheses given in (3.18a) and (3.18b) matches the definition in (3.5) for the special case where the dictionary contains just one spectrum. In this special case, the spectral input \( f^* \) is in the dictionary under the null hypothesis in both DSD and ASD problem formulations.\(^2\)

3.5.2 Anomaly detection approach

Our anomaly detection approach and the associated theoretical analysis are based on a “distance preservation” property of \( A \), which is stated formally in (3.20). We propose an anomaly detection method that controls the false discovery rate (FDR) below a desired level \( \delta \) for different background and sensor noise statistics. In other

---

\(^1\) Note that \( \tau \) cannot exceed \( \sqrt{2} \) because we assume that all targets of interest, including those in \( \mathcal{D} \) and the actual target \( f^* \), are unit-norm.

\(^2\) The anomaly detection problem discussed here is more accurately described as target detection in the classical detection theory vocabulary. However, in recent works [81, 82], the authors assume that the nominal distribution is obtained from training data and a test sample is declared to be anomalous if it falls outside of the nominal distribution learned from the training data. Our work is in a similar spirit where we learn our spectral dictionary from training data and label any test spectrum that does not correspond to our spectral dictionary as being anomalous.
Figure 3.1: Illustration of the anomaly detection hypothesis tests. Consider that \( \mathcal{D} = \{ f^{(1)}, \ldots, f^{(4)} \} \). A unit-norm target spectrum \( f^* \in \mathbb{R}^N \) is considered anomalous if it falls in the shaded region.

In words, we control the expected ratio of falsely declared anomalies to the total number of spectra declared to be anomalous. Note that here we work with the FDR as opposed to the pFDR, since it is possible for a scene to not contain any anomalies at all. We let \( V/R = 0 \) for \( R = V = 0 \) since one does not declare any spectrum to be anomalous in this case. Our goal is to develop a detection approach that controls the FDR below a desired level \( \delta \). In [37], Benjamini and Hochberg discuss a p-value based procedure, “BH procedure”, that controls the false discovery rate of \( M \) independent hypothesis tests below a desired level. Let

\[
d_i = \min_{f \in \mathcal{D}} \| y_i - \alpha_i A f \| = \min_{f \in \mathcal{D}} \| \alpha_i A (f^*_i - f) + n_i \|
\]

be the test statistic at the \( i^{th} \) spatial location. The p-value can be defined in terms of our test statistic as follows:

\[
p_i = \mathbb{P}(\tilde{d}_i \geq d_i | \mathcal{H}_{0i}) \quad (3.19)
\]

where \( \tilde{d}_i = \min_{f \in \mathcal{D}} \| \alpha_i A (f^*_i - f) + n_i \| \) and \( n \sim \mathcal{N}(0, I) \) is independent of \( n_i \). This is the probability under the null hypothesis, of acquiring a test statistic at least as extreme as the one observed. Let us denote the ordered set of p-values by \( p(1) \leq p(2) \leq \cdots \leq p(M) \) and let \( \mathcal{H}_{0i} \) be the null hypothesis corresponding to \( i^{th} \) p-value. The BH procedure says that if we reject all \( \mathcal{H}_{0i} \) for \( i = 1, \ldots, t \) where \( t \) is the largest \( i \) for which \( p(i) \leq i\delta/M \), then the FDR is controlled at \( \delta \).
To apply this procedure in our setting, we need to find a tractable expression for the p-value at every spatial location. This can be accomplished when \( A \) satisfies the distance-preservation condition stated below. Let \( V = D \cup \{ f_i^* : i \in \{1, \ldots, M\} \} \) be the set of all spectra in the dictionary and spectral image. Note that \(|V| \leq M + m\).

For a given \( \epsilon \in (0, 1) \), a projection operator \( A \in \mathbb{R}^{K \times N} \), \( K \leq N \), is distance-preserving on \( V \) if the following holds for all \( u, v \in V \):

\[
(1 - \epsilon)\|u - v\| \leq \|A(u - v)\| \leq (1 + \epsilon)\|u - v\|, \forall u, v \in V.
\]

The existence of such projection operators is guaranteed by the celebrated Johnson and Lindenstrauss (JL) lemma [103], which says that there exists random constructions of \( A \) for which (3.20) holds with probability at least \( 1 - 2|V|^2e^{-Kc(\epsilon)} \) provided \( K = O(\log |V|) \leq N \), where \( c(\epsilon) = \epsilon^2/16 - \epsilon^3/48 \) [38, 39]. Examples of such constructions are: (a) Gaussian matrices whose entries are drawn from \( \mathcal{N}(0, 1/K) \), (b) Bernoulli matrices whose entries are \( \pm 1/\sqrt{N} \) with probability \( 1/2 \), (c) random matrices whose entries are \( \pm \sqrt{3/N} \) with probability \( 1/6 \) and zero with probability \( 2/3 \) [38, 39], and (d) matrices that satisfy the Restricted Isometry Property (RIP) where the signs of the entries in each column are randomized [40].

We now state our main theorem that gives a tight upper bound on the p-value at every spatial location when \( \{\alpha_i\} \) are unknown and are estimated from the observations. Let \( \{\hat{\alpha}_i\} \) be the estimates of \( \{\alpha_i\} \) that satisfy

\[
1 - \zeta \leq \frac{\alpha_i}{\hat{\alpha}_i} \leq 1 + \zeta
\]

for \( i = 1, \ldots, M \) where \( \zeta \in [0, 1] \) is a measure of the accuracy of the estimation procedure.

**Theorem 7.** If the \( i^{th} \) hypothesis test is defined according to (3.18a) and (3.18b), the projection matrix \( A \) satisfies (3.20) for a given \( \epsilon \in (0, 1) \), and the estimates \( \{\hat{\alpha}_i\} \) satisfy (3.21) for some \( \zeta \in [0, 1] \), then the bound

\[
p_i \leq 1 - \mathcal{F} \left( d_i^2; K, (1 + \epsilon)^2\hat{\alpha}_i^2 (\zeta + \tau)^2 \right)
\]

(3.22)
holds for all $i = 1, \ldots, M$ where $F(\cdot; K, \nu)$ is the CDF of a noncentral $\chi^2$ random variable with $K$ degrees of freedom and noncentrality parameter $\nu$ [105].

The proof of this theorem is given in Appendix A.2.5. We find the p-value upper bounds at every spatial location and use the BH procedure to perform anomaly detection. The performance of this procedure depends on the values of $K$, $\{\alpha_i\}$, $\tau$ and $\epsilon$. The parameter $\epsilon$ is a measure of the accuracy with which the projection matrix $A$ preserves the distances between any two vectors in $\mathbb{R}^N$. A value of $\epsilon$ close to zero implies that the distances are preserved fairly accurately. When $\{\alpha_i\}$ are unknown and estimated from the observations, the performance depends on the accuracy of the estimation procedure, which is reflected in our bounds in (3.22) through $\zeta$.

One can easily estimate $\{\alpha_i\}$ from $\{y_i\}$ for some choices of $A$. For instance, if the entries of the projection matrix $A$ are drawn from $\mathcal{N}(0, 1/K)$, the $\{\alpha_i\}$ can be estimated using a maximum likelihood estimator (MLE), which in this case is given by

$$\hat{\alpha}_i = \sqrt{||y_i||^2 - K}. \quad (3.23)$$

The experimental results discussed in Sec. 3.6 demonstrate the performance of this detector as a function of $K$, $\{\alpha_i\}$ and $\tau$ when $\{\alpha_i\}$ are known and as a function of $K$, $\tau$ and $\zeta$ when $\{\alpha_i\}$ are estimated.

### 3.6 Experimental Results

In the experiments that follow, the entries of $A$ are drawn from $\mathcal{N}(0, 1/K)$.

#### 3.6.1 Dictionary spectra detection

To test the effectiveness of our approach, we formed a dictionary $\mathcal{D}$ of nine spectra (corresponding to different kinds of trees, grass, water bodies and roads) obtained from a labeled HyMap (Hyperspectral Mapper) remote sensing data set [2], and
simulated a realistic dataset using the spectra from this dictionary. Each HyMap spectrum is of length \(N = 106\). We generated projection measurements of these data such that \(z_i = \alpha_i \Phi(f_i^* + b_i) + w_i\) according to (3.1), where \(w_i \sim \mathcal{N}(0, \sigma^2 I)\), \(f_i^* \in \mathcal{D}\) for \(i = 1, \ldots, 8100\), \(b_i \sim \mathcal{N}(\mu_b, \Sigma_b)\) such that \(\Sigma_b\) satisfies the condition in (3.4), and \(\alpha_i = \alpha_i^* \sqrt{K}\) where \(\alpha_i^* \sim \mathcal{U}[21, 25]\) and \(\mathcal{U}\) denotes uniform distribution. We let \(\sigma^2 = 5\) and model \(\{\alpha_i\}\) to be proportional to \(\sqrt{K}\) to account for the fact that the total observed signal energy increases as the number of detectors increases. We transform the \(z_i\) by a series of operations to arrive at a model of the form discussed in (3.2), which is \(y_i = \alpha_i A f_i^* + n_i\). For this dataset, \(p_{\min} = 0.04938\), \(p_{\max} = 0.1481\), and \(d_{\min} = 0.04341\).

We evaluate the performance of our detector (3.7) on the transformed observations, relative to the number of measurements \(K\), by comparing the detection results to the ground truth. Our MAP detector returns a label \(L_i^{\text{MAP}}\) for every observed spectrum which is determined according to

\[
L_i^{\text{MAP}} = \arg \min_{\ell \in \{1, \ldots, m\}, f^{(\ell)} \in \mathcal{D}} \left( \frac{1}{2} \|y_i - \alpha_i A f^{(\ell)}\|^2 - \log p^{(\ell)} \right)
\]

where \(m\) is the number of spectra in \(\mathcal{D}\), and \(p^{(\ell)}\) is the a priori probability of target class \(\ell\). In our experiments we evaluate the performance of our classifier when (a) \(\{\alpha_i\}\) are known (AK) and (b) \(\{\alpha_i\}\) are unknown (AU) and must be estimated from \(y\), respectively. The empirical pFDR\(^{(j)}\) for each target spectrum \(j\) is calculated as follows:

\[
pFDR^{(j)} = \frac{\sum_{i=1}^M I\{L_{i}^{\text{GT}} = j\} I\{L_{i}^{\text{MAP}} \neq j\}}{\sum_{i=1}^M I\{L_{i}^{\text{MAP}} \neq j\}}
\]

where \(\{L_{i}^{\text{GT}}\}\) denote the ground truth labels. The empirical pFDR\(^{(j)}\) is the ratio of the number of missed targets to the total number of spectra that were declared to
be nontargets. The plots in Fig. 3.2(a) show the results obtained using our target detection approach under the AK case (shown by a dark gray dashed line) and the AU case (shown by a light gray dashed line), compared to the theoretical upper bound (shown by a solid line). These results are obtained by averaging the pFDR values obtained over 1000 different noise, sensing matrix and background realizations. Note that theoretical results only apply to the AK case since they were derived under the assumption of \( \{\alpha_i\} \) being known. The experimental results are shown for both AK and AU cases to provide a comparison between the two scenarios. In both these cases, the worst-case empirical pFDR curves decay with the increase in the values of \( K \). In the AK case, in particular, the worst-case empirical pFDR curve decays at the same rate as the upper bound. In this experiment, for a fixed \( \alpha_{\min} \) and \( d_{\min} \), we chose \( K \) to satisfy (3.17). The theory is somewhat conservative, and in practice the method works well even when the values of \( K \) are below the bound in (3.17).

Figure 3.2: Compressive target detection results under the AK and AU cases respectively as a function of \( K \).

In the experiment that follows, we let \( \alpha_i^* \sim U[10, 20] \), where \( U \) denotes a uniform random variable, \( \alpha_i = \sqrt{K} \alpha_i^* \) and evaluate the performance of our detector for different values of \( K \) that are not necessarily chosen to satisfy (3.17). In addition, we also compare the performance of our detection method to that of a nearest-neighbor
(NN) based target detector operating on multispectral versions of our simulated spectral input image.

For an input spectrum \( g \in \mathbb{R}^N \), we let \( \tilde{g} \in \mathbb{R}^K \) denote its multispectral approximation. Specifically, the \( j \)th element of \( \tilde{g} \) is \( \sum_{\ell=1}^{r} g_{(j-1)r+\ell} \) where \( r = \lceil N/K \rceil \). (We also tried forming multispectral versions of \( g \) using point sampling and achieved similar, unreported, results.) For the multispectral NN-based detection method, let us consider making observations of the form

\[
y_i = \frac{\tilde{g}_i}{c} + n_i \in \mathbb{R}^K
\]

(3.24)

where \( \tilde{g}_i = \alpha_i \tilde{f}_i^* + \tilde{b}_i \) is the \( K \)-dimensional multispectral version of \( f_i^* + b_i \) for \( K \leq N \), \( n_i \sim \mathcal{N}(0, \sigma^2 I) \) for \( \sigma^2 = 5 \) and \( c \) is a constant that is chosen to preserve the mean signal-to-noise ratio corresponding to the multispectral and projection measurements. The multispectral NN-based detector returns a label \( L_i^{NN} \) for every observed spectrum which is determined according to

\[
L_i^{NN} = \arg \min_{\ell \in \{1, \ldots, m\}, f^{(\ell)} \in \mathcal{D}} \left( y_i - \alpha_i \tilde{f}^{(\ell)} \right)^T G^{-1} \left( y_i - \alpha_i \tilde{f}^{(\ell)} \right)
\]

where \( G = \tilde{\Sigma}_b + \sigma^2 I \) and \( \tilde{\Sigma}_b \) is the covariance matrix obtained from the multispectral versions of the background training data and \( \tilde{f}^{(\ell)} \) is the multispectral version of \( f^{(\ell)} \in \mathcal{D} \). The algorithm declares that target spectrum \( f^{(j)} \in \mathcal{D} \) is present in the \( i \)th spatial location if \( L_i^{NN} = j \). Fig. 3.2(b) shows a comparison of the results obtained using our detection method and the multispectral NN-based detection method under the AK case. These results show that our methods performs significantly better than the alternate method, and that the empirical pFDR values in our method decays with \( K \).
3.6.2 Anomaly detection

In this section, we evaluate the performance of our anomaly detection method on (a) a simulated dataset and provide a comparison with the performance of a classical likelihood ratio test (LRT) based anomaly detection approach on a multispectral version of our spectral input image, and (b) real AVIRIS (Airborne Visible InfraRed Imaging Spectrometer) dataset.

Experiments on simulated data

We simulate a spectral image $f^*$ composed of 8100 spectra, where each of them is either drawn from a spectral dictionary $D = \{f^{(1)}, \cdots, f^{(5)}\}$ consisting of five labeled spectra from the HyMap data that correspond to a natural landscape (trees, grass and lakes) or is anomalous. The anomalous spectrum is extracted from unlabeled AVIRIS data, and the minimum distance between the anomalous spectrum $f^{(a)}$ and any of the spectra in $D$ is $d_{\min} = \min_{f \in D} \|f - f^{(a)}\| = 0.5308$. The simulated data has 625 locations that contain the anomalous spectrum. Our goal is to find the spatial locations that contain the anomalous AVIRIS spectrum given noisy measurements of the form $z_i = \Phi (\alpha_i f_i^* + b_i) + w_i$ where $b_i \sim (\mu_b, \Sigma_b)$, $\Phi$ is designed according to (A.10), $w_i \sim N(0, \sigma^2 I)$ and $f_i^* \in D$ under $\mathcal{H}_{0i}$. As discussed in Sec. 3.5, $f_i^*$ is anomalous under $\mathcal{H}_{1i}$, and our goal is to control the FDR below a user-specified false discovery level $\delta$. We simulate $\{\alpha_i\} = \sqrt{K} \alpha^*_i$ where $\alpha^*_i \sim U[2, 3]$. In this experiment we assume the availability of background training data to estimate the background statistics and the sensor noise variance $\sigma^2$. Given the knowledge of the background statistics, we perform the whitening transformation discussed in Sec. 3.2 and evaluate the detection performance on the preprocessed observations given by (3.2).

For a fixed $\tau = 0.1$ and $\epsilon = 0.1$, we evaluate the performance of the detector as the number of measurements $K$ increases under the AK and AU cases respectively, by comparing the pseudo-ROC (receiver operating characteristic) curves obtained by
plotting the empirical false discovery rate against the probability of correct rejection 
(1 − FNR, where FNR is the false nondiscovery rate). The empirical FDR and FNR 
are computed according to

\[
\text{FDR} = \frac{\sum_{i=1}^{M} \mathbb{I}\{L_{i}^{GT}=0\} \mathbb{I}\{p_i \leq p_t\}}{\mathbb{I}\{p_i \leq p_t\}}
\]

\[
\text{FNR} = \frac{\sum_{i=1}^{M} \mathbb{I}\{L_{i}^{GT}=1\} \mathbb{I}\{p_i > p_t\}}{\mathbb{I}\{p_i > p_t\}}
\]

where \(p_t\) is the p-value threshold such that the BH procedure rejects all null hypotheses for which \(p_i \leq p_t\), and the ground truth label \(L_{i}^{GT} = 0\) if the \(i^{th}\) spectrum is not anomalous, and 1 otherwise. In this experiment, we consider four different values of 
\(K\) approximately given by \(K \in \{N/6, N/3, N/2\}\) where \(N = 106\), and evaluate the 
performance of our detector for each \(K\).

To compare the performance of our method to a classical detection method, let 
us consider an LRT-based procedure, where we collect measurements of the form in 
(3.24) and \(f_i^* \in D\) under \(H_0\). Observe that \(y_i|H_0 \sim \sum f_i \in D \mathbb{P}(f_i = f) \mathcal{N}(\alpha_i \tilde{f}, \tilde{\Sigma}_b + I)\), where \(\tilde{f}\) refers to the multispectral version of \(f \in D\). In this experiment we 
assume that each spectrum in \(D\) is equally likely under \(H_0\) for \(i = 1, \ldots, M\). The 
GLRT-based approach declares the \(i^{th}\) spectrum to be anomalous if

\[
- \log \mathbb{P}(y_i|H_0) \overset{H_1}{\gtrless} \eta
\]

for \(i = 1, \ldots, M\), where \(\eta\) is a user-specified threshold [83]. While our anomaly 
detection method is designed to control the FDR below a user-specified threshold, 
the GLRT-based method is designed to increase the probability of detection while keeping the probability of false alarm as low as possible. To facilitate a fair evaluation 
of these methods, we compare the pseudo-ROC curves (FDR versus 1-FNR) and the 
actual ROC curves (probability of false alarm \(p_f\) versus probability of detection \(p_d\))
corresponding to these methods obtained by averaging the empirical FDR, FNR, \( p_d \) and \( p_f \) over 1000 different noise and sensing matrix realizations for different values of \( K \). Figs. 3.3(a) and 3.3(b) show the pseudo-ROC plots and the conventional ROC plots obtained using the multispectral GLRT-based method when \( \{\alpha_i\} \) are known. Figs. 3.3(c) and 3.3(d) show the pseudo-ROC plots and the conventional ROC plots obtained using our method when \( \{\alpha_i\} \) are known. These plots show that performing anomaly detection from projection measurements yields better results than performing anomaly detection on multispectral measurements. This is largely due to the fact that carefully chosen projection measurements preserve distances (up to a constant factor) among pairs of vectors in a finite collection, where as the multispectral measurements fail to preserve distances among vectors that are very similar to each other. Figs. 3.3(e) and 3.3(f) show the pseudo-ROC plots and the conventional ROC plots obtained using our method when \( \{\alpha_i\} \) are unknown, and are estimated from the measurements. Note that the value of \( \zeta \) decreases as \( K \) increases since the estimation accuracy of \( \{\alpha_i\} \) increases with increase in \( K \). These plots show that the performance improves as we collect more observations, and that, as expected, the performance under the AK case is better than the performance under the AU case.

**Experiments on real AVIRIS data**

To test the performance of our anomaly detector on a real dataset, we consider the unlabeled AVIRIS Jasper Ridge dataset \( \mathbf{g} \in \mathbb{R}^{614 \times 512 \times 197} \), which is publicly available from the NASA AVIRIS website, [http://aviris.jpl.nasa.gov/html/aviris.freedata.html](http://aviris.jpl.nasa.gov/html/aviris.freedata.html). We split this data spatially to form equisized training and validation datasets, \( \mathbf{g}^t \) and \( \mathbf{g}^v \) respectively, each of which is of size \( 128 \times 128 \times 197 \). Figs. 3.4(a) and 3.4(b) show images of the AVIRIS training and validation data summed through the spectral coordinates. The training data are comprised of a
rocky terrain with a small patch of trees. The validation data seems to be made of a similar rocky terrain, but also contain an anomalous lake-like structure. The goal is to evaluate the performance of the detector in detecting the anomalous region in the validation data for different values of $K$. We cluster the spectral targets in the normalized training data to eight different clusters using the K-means clustering algorithm and form a spectral dictionary $D$ comprising of the cluster centroids. Given the spectral dictionary and the validation data, we find the ground truth by labeling the $i^{th}$ validation spectrum as anomalous if $\min_{f \in D} \left\| f - \frac{g^v}{\|g^v\|} \right\| > \tau$. Since the statistics of the possible background contamination in the data could not be learned in this experiment because of the lack of labeled training data, the spectral dictionary might be background contaminated as well. The parameter $\tau$ encapsulates this uncertainty in our knowledge of the spectral dictionary. In this experiment, we set $\tau = 0.2$.

We generate measurements of the form $y_i = \sqrt{K}g^v_i + n_i$ for $i = 1, \ldots, 128 \times 128$, where $n_i \sim N(0, I)$. The $\sqrt{K}$ factor indicates that the observed signal strength increases with $K$. For a fixed FDR control value of 0.01, Figs. 3.4(c) and 3.4(d) show the results obtained for $K \approx N/5$ and $K \approx N/2$ respectively. Fig. 3.4(e) shows how the probability of error decays as a function of the number of measurements $K$. The results presented here are obtained by averaging over 1000 different noise and sensing matrix realizations. From these results, we can see that the number of detected anomalies increases with $K$ and the number of misclassifications decrease with $K$.

### 3.7 Conclusion

This work presents computationally efficient approaches for detecting spectral targets and anomalies of different strengths from spectral projection measurements without
performing a complete reconstruction of the spectral image, and offers theoretical bounds on the worst-case target detector performance. This work treats each spectrum in the spectral image as independent of its spatial neighbors. This assumption is reasonable in many contexts, especially when the spatial resolution is low relative to the spatial homogeneity of the environment. However, emerging technologies in computational optical systems continue to improve the resolution of spectral imagers. In our future work we will build upon the methods that we have discussed here to exploit the spatial correlations in the data.
Figure 3.3: Comparison of the performances of our anomaly detector and the multispectral GLRT-based detector for different values of $K$ when $\alpha_i^* \in U[2,3]$ and $\alpha_i = \alpha_i^* \sqrt{K}$. Note the different scales on the axes.
Figure 3.4: Anomaly detection results corresponding to real AVIRIS data for a fixed FDR control of 0.01.
Fast level set estimation from projection measurements

Estimation of the level set of a function (i.e., regions where the function exceeds some value) is an important problem with applications in digital elevation mapping, medical imaging, and astronomy. In many applications, however, the function of interest is acquired through indirect measurements, such as tomographic projections, coded-aperture measurements, or pseudo-random projections associated with compressed sensing. This chapter describes a new methodology and associated theoretical analysis for rapid and accurate estimation of the level set from such projection measurements. The proposed method estimates the level set from projection measurements without an intermediate function reconstruction step, thereby leading to significantly faster computation. In addition, the coherence of the projection operator and McDiarmid’s inequality are used to characterize the estimator’s performance.

4.1 Level set estimation

Level set estimation is the process of using observations of a function \( f \) defined on the unit square \([0, 1]^2\) to estimate the region(s) where \( f \) exceeds some critical value \( \gamma \);
i.e. \( S^* \equiv \{ x \in [0,1]^2 : f(x) \geq \gamma \} \). Accurate and efficient level set estimation plays a crucial role in a variety of scientific and engineering tasks, including the localization of “hot spots” signifying tumors in medical imaging, significant photon sources in astronomy, or strong reflectors in remote sensing. In this work, we consider a discrete function \( f \in \mathbb{R}^{N_1 \times N_2} \), which is obtained from \( f \) via integration sampling, i.e.,

\[
 f_{i_1, i_2} = \int_{i_1/N}^{(i_1+1)/N} \int_{i_2/N}^{(i_2+1)/N} f(x_1, x_2) dx_1 dx_2
\]

for \( i_1 = 0, 1, \ldots, N_1 - 1 \) and \( i_2 = 0, 1, \ldots, N_2 - 1 \). Let \( f \in \mathbb{R}^N \), \( N = N_1 \times N_2 \) denote the vectorized function \( f \) and \( f_i \) represent the \( i \)th element of \( f \). A \( \gamma \)-level set in this discrete setting can be written as \( S^* = \{ i : f_i > \gamma \} \).

Previous work [44] explored the estimation of level sets of a function \( f \) from noisy observations of the form \( \tilde{y} = f + n \), where \( n \) denotes a vector of independent, zero-mean noise realizations. However, there are many contexts where direct observations of this form are not available; instead, we make observations of the form \( y = Af + n \), where \( A \) is a discrete linear projection operator that may not be invertible. For instance, \( y \) might correspond to tomographic projections in tomography, multiple blurred, low-resolution, dithered snapshots in astronomy, or pseudo-random projections in compressed sensing systems [106].

Our goal in this setting is to perform level set estimation without an intermediate step involving time-consuming reconstruction of \( f \). There are two reasons for this: First, level set estimation without reconstruction of \( f \) would allow sequential measurement schemes to be performed on the fly. For instance, in tomography we would like to estimate \( S^* \) quickly from the observations so that additional data focused on \( S^* \) can be collected immediately, resulting in an overall low radiation dose. Second, “plug-in” approaches that estimate \( f \) and threshold the estimate \( \hat{f} \) to extract \( S^* \) are notoriously difficult to characterize; the performance hinges upon the statistics of the function estimation error \( \hat{f} - f \), which for most reconstruction methods are
unknown (with the possible exception of the first moment). More generally, reconstruction methods aim to minimize the total error, integrated or averaged spatially over the entire function. This does little to control the error at specific locations of interest, such as in the vicinity of the level set boundary.

4.1.1 Problem formulation

We are interested in estimating a $\gamma$ level set $S^* = \{i : f_i > \gamma\}$ of $f$ from projection measurements of the form $y = Af + n \in \mathbb{R}^K$ for $K \leq N$, where $A \in \mathbb{R}^{K \times N}$ is a projection operator that is assumed to be known, and $n$ is bounded, independent and zero-mean noise $^1$. Our main goal is to estimate $S^*$ from $y$ without reconstructing the underlying function $f$.

4.2 Our contribution and relation with previous work

In this work, we demonstrate that, subject to certain conditions on the (linear) projection operator $A$ and the $\ell_1$ norm of $f$, the level set $S^*$ can be estimated quickly and accurately without first reconstructing $f$. For $A = I$, [44] provides minimax optimal, tree-based level set estimation techniques to extract $S^*$ from noisy observations $\tilde{y} = f + n \in \mathbb{R}^N$ without estimating $f$. We cannot directly apply the strategy in [44] to our problem since $A \neq I$. However, we draw on the key idea of constructing proxy observations

$$z = A^T y = f + \left( A^T A - I \right) f + A^T n. \quad (4.1)$$

from the sparse support detection literature [41–43] and then exploit some of the important insights from [44] to address our problem.

Before we present our estimation method, we discuss prior works on level set estimation and sparse support detection.

$^1$ This assumption is reasonable, since, in practice, the noise is always bounded due to hardware limitations in physical sensors.
4.2.1 Previous work on level set estimation

The basic idea in [44] was to design an estimator of the form \( \hat{S} = \arg\min_{S \in \mathcal{S}} \hat{R}(S) + \Phi(S) \), where \( \mathcal{S} \) is a class of candidate estimates, \( \hat{R} \) is an empirical measure of the estimator risk based on \( N \) noisy observations of the function \( f \), and \( \Phi \) is a regularization term which penalizes improbable level sets. They described choices for \( \hat{R}, \Phi, \) and \( \mathcal{S} \) which made \( \hat{S} \) rapidly computable and minimax optimal for a large class of level set problems. In particular, [44] proposed a novel error metric between \( S^* \) and a candidate estimate \( S \) that was ideally suited to the problem at hand. That error metric in a discrete setting can be written as

\[
\varepsilon(S, S^*) = \frac{1}{N} \sum_{i \in \Delta(S^*, S)} |\gamma - f_i| \tag{4.2}
\]

where \( \Delta(S^*, S) \equiv \{ i \in (S^* \cap S) \cup (S^* \cap S) \} \) denotes the symmetric difference, and \( S \) is the complement of \( S \). Note that (4.2) is simply a weighted probability of error where the weights depend on the amplitude of the function relative to the level set threshold \( \gamma \). While the expression in (4.2) is not directly computable (since \( f \) and thus \( S^* \) are unknown), one can nevertheless minimize it by defining the following risk function:

\[
R(S) \equiv \frac{1}{N} \sum_i \ell_i(S) \tag{4.3}
\]

where

\[
\ell_i(S) = (\gamma - f_i) [\mathbb{I}_{\{i \in S\}} - \mathbb{I}_{\{i \notin S\}}] \tag{4.4}
\]

is the loss function and \( \mathbb{I}_{\{E\}} = 1 \) if event \( E \) is true and 0 otherwise. The loss function in (4.4) measures the distance between the function at location \( i \), \( f_i \), and the threshold, \( \gamma \), and weights this distance by \( \pm 1 \) according to whether \( i \in S \) or not. The loss function \( \ell_i(S) \) is positive if \( i \in \Delta(S^*, S) \) and is negative otherwise. To
see this, observe that for all $i \in S^* \cap \Sigma$, $(\gamma - f_i) \leq 0$ and $[\mathbb{I}_{i \in \mathcal{S}} - \mathbb{I}_{i \in \mathcal{S}}] = -1$. A similar explanation holds for all $i \in \Sigma^c \cap S$ as well. Note that $R(S) - R(S^*)$ is equivalent to the error metric $\varepsilon_{(S,S^*)}$ defined in (4.2),

$$R(S) - R(S^*) = \frac{1}{N} \sum_i (\gamma - f_i) \left( [\mathbb{I}_{i \in \mathcal{S}} - \mathbb{I}_{i \in \mathcal{S}}] - [\mathbb{I}_{i \in \mathcal{S}^*} - \mathbb{I}_{i \in \mathcal{S}^*}] \right)$$

$$= \frac{2}{N} \sum_{i \in \Delta(S^*, S)} |\gamma - f_i| = 2\varepsilon_{(S,S^*)}.$$

Finding an estimator that minimizes the error $\varepsilon_{(S,S^*)}$ is equivalent to finding an estimator that minimizes $R(S)$ since $R(S^*)$ is simply a constant. Furthermore, [44] showed that it is possible to estimate $R(S)$ from $\tilde{y}$ in a straightforward manner.

Using Hoeffding’s inequality, [44] derived a regularization term $\Phi$ and developed a dyadic tree-based framework which can be used to obtain $\hat{S}$. Trees were utilized for a couple of reasons. First, they both restricted and structured the space of potential estimators in a way that allowed the global optimum to be both rapidly computable and very close to the best possible (not necessarily tree-based) estimator. Second, they allowed the estimator selection criterion to be spatially adaptive, which is critical for the formation of provably optimal estimators. The presence of a non-unitary $A$ and correlated noise $n'$ in our problem setup make a direct extension of the techniques in [44] to our problem framework nontrivial.

4.2.2 Relationship with previous work on sparse support detection

The idea of constructing proxy observations $z$ to deduce certain properties of the underlying $f$ has been successfully employed in recent compressed sensing and statistics literature to solve the problem of support detection of a discrete $f$ having no more than $m$ non-zero entries; see, e.g., [41–43]. In fact, our level set estimation methodology is inspired by the empirical and theoretical success of using thresholded proxy observations for support detection of (discrete) sparse functions. However, despite

76
the fact that both our method and the thresholded support detection of [41–43] make use of a function proxy, there are some key differences between the two lines of work that stem partly from different underlying assumptions. Specifically, it is established in [41] that the support of an $m$-sparse $f$ can be reliably detected from appropriately thresholded proxy observations with an overwhelming probability as long as $A$ satisfies a certain coherence property. On the other hand, the level set estimation analysis carried out in this work does not impose any sparsity constraints on the underlying function $f$. Indeed, it is not difficult to convince oneself that directly thresholding the proxy observations for level set estimation in the case of a non-sparse $f$ would lead to numerous false positives and false negatives (see, e.g., Figs. 4.2(c-d), 4.4(c-d), 4.6(c-d), 4.8(c-d), 4.10(c-d), 4.12(c-d), 4.14(c-d), 4.16(c-d), 4.18(c-d), 4.20(c-d), 4.22(c-d), 4.24(c-d), 4.26(c-d), 4.28(c-d), 4.30(c-d)). In contrast, our methodology relies on a novel two-step approach that enables us to work with proxy observations without requiring $f$ to be strictly sparse. In our experiments, the proposed method performs an order of magnitude better than sparse support estimation methods in [41–43] with an appropriate threshold.

4.2.3 Our contribution

There are two key challenges we address in our analysis to specify the proposed estimator’s performance. First, we must characterize $n'$, which can be considered a combination of noise and interference caused by calculations of the proxy observations; the interference plays a crucial role in our estimation error and scales with the worst-case coherence of the projection operator $A$ and the $\ell_1$ norm of $f$. Second, the original analysis in [44] considered independent noise realizations, allowing for the application of Hoeffding’s inequality to analyze estimator’s performance. However, $n'$ in our proxy observations contains statistical dependencies which we consider in our revised analysis. Our theoretical analysis is validated via a medical imaging
4.3 Fast level set estimation from projection measurements

In order to extract the $\gamma$-level set of $f$ from $y$, we propose a novel two-step procedure. First, we construct a proxy of $f$ according to (4.1) since this allows us to arrive at the canonical signal plus noise observation model. Next, we perform level set estimation on the proxy observations $z$, rather than on $y$, by relying on the insights of [44]. Note that for any unitary $A$, $z$ in (4.1) reduces to $\tilde{y}$. However, for non-unitary $A$, the proxy defined in (4.1) creates a signal-dependent interference term $(A^T A - I) f$ and a zero-mean correlated noise term $A^T n$. This interference term $n'$ makes a direct extension of the level set approach discussed in [44] to our problem nontrivial. In our work, we theoretically analyze the impact of $n'$ and use the theoretical results to develop a dyadic, tree-based level set estimation approach that adapts to the interference term.

Given $z$, our goal is to find a level set estimate $\tilde{S} = \arg \min_{S \in S} R(S) - R(S^*)$ where $S$ is a family of candidate level set estimates and $R(\cdot)$ is defined in (4.3). (Note that $\tilde{S} = S^*$ if $S^* \in S$.) Since $f$ is unknown, we consider an empirical risk of the form

$$\hat{R}(S) = \frac{1}{N} \sum_{i=1}^{N} (\gamma - z_i) \left[ I_{\{i \in S\}} - I_{\{i \notin S\}} \right]$$

(4.5)

and show that finding an estimate $\hat{S} = \arg \min_{S \in S} \hat{R}(S) + \Phi(S)$, where $\Phi(S)$ is an interference-dependent penalty term, yields $\left| R(\hat{S}) - R(\tilde{S}) \right| \overset{N}{\rightarrow} 0$. The penalty term plays a major role in our estimation strategy and is crucial in finding estimates that hone in on the boundary of the level set $S^*$. We thus focus on designing a spatially adaptive penalty $\Phi(S)$ that promotes well-localized level sets with potentially non-smooth boundaries. Following the analysis in [44] we let $S$ be a family of level set
estimates defined on recursive dyadic partitions of the domain of \( f \); e.g., an image could be partitioned into patches of varying side-lengths using a quad-tree, so that each leaf of the tree corresponded to one patch. Each leaf in the partition would be estimated to be in or out of the level set of interest. Let \( \pi(S) \) be the partition induced by an estimate \( S \in \mathcal{S} \). Then, the risk of \( S \) in each of its leaf \( L \in \pi(S) \) is given by

\[
R(L) = \frac{1}{N} \sum_{i=1}^{N} \left( \gamma - f_i \right) \left[ I_{\{\ell(L)=1\}} - I_{\{\ell(L)=0\}} \right] I_{\{i \in L\}}
\]

where \( \ell(L) = 1 \) if \( L \) is in the level set and 0 otherwise. Note that \( R(S) = \sum_{L \in \pi(S)} R(L) \). We design a spatially adaptive penalty term by analyzing \( R(L) - \hat{R}(L) \) at each leaf separately. To facilitate our analysis, let us define

\[
\tilde{R}(L) = \frac{1}{N} \sum_{i=1}^{N} \left( \gamma - \mathbb{E}[z_i] \right) \left[ I_{\{\ell(L)=1\}} - I_{\{\ell(L)=0\}} \right] I_{\{i \in L\}}
\]

Then

\[
R(L) - \hat{R}(L) = R(L) - \tilde{R}(L) + \tilde{R}(L) - \hat{R}(L)
\]

\[
\equiv \frac{1}{N} \sum_{i=1}^{N} \left( \mathbb{E}[z_i] - f_i \right) \left[ I_{\{\ell(L)=1\}} - I_{\{\ell(L)=0\}} \right] I_{\{i \in L\}}
\]

\[
+ \frac{1}{N} \sum_{i=1}^{N} (z_i - \mathbb{E}[z_i]) \left[ I_{\{\ell(L)=1\}} - I_{\{\ell(L)=0\}} \right] I_{\{i \in L\}}.
\]  

(4.6)

Note that while \( T_1 \) is a measure of the biases of \( \{z_i\} \), \( T_2 \) is a measure of the concentration of \( \{z_i\} \) around their means. Let us consider the statistics of \( z \) to further understand \( T_1 \) and \( T_2 \) respectively. Assuming without loss of generality that the columns of \( A \) have unit \( \ell_2 \) norms, one can easily see from (4.1) that

\[
z_i = f_i + \sum_{j=1,j \neq i}^{N} f_j \langle A^{(i)}, A^{(j)} \rangle + \langle A^{(i)}, n \rangle
\]

where \( A^{(i)} \) denotes the \( i \)th column of \( A \). Since \( A \) is given, and \( n \) is zero mean, the term

\[
\mathbb{E}[z_i] - f_i = \sum_{j=1,j \neq i}^{N} f_j \langle A^{(i)}, A^{(j)} \rangle
\]

79
in $T_1$ is the signal-dependent interference term at the $i^{th}$ location due to the signal energies at other locations. We upper bound $T_1$ by the $\ell_1$ norm of $f$ and the worst-case coherence of $A$, and bound $T_2$ using McDiarmid’s inequality [107] and sum the risk in each leaf of the estimate $S$ to arrive at our main result stated below:

**Theorem 8.** Suppose that the entries of noise $n$ are bounded between $[c_\ell, c_u]$. Then, for $\delta \in [0, 1/2]$, with probability at least $1 - \delta$, the following holds for all $S \in S$:

$$R(S) \leq \hat{R}(S) + \left( \frac{N - 1}{N} \right) \mu(A) \|f\|_1 +$$

$$\sum_{L \in \pi(S)} \frac{1}{N} \left[ \log(2/\delta) + [L] \log 2 \right] |c_u - c_\ell|^2 \sum_{i,j \in L} \langle A^{(i)}, A^{(j)} \rangle$$

where $\mu(A) = \max_{i,j \in \{1,\ldots,N\}, i \neq j} |\langle A^{(i)}, A^{(j)} \rangle|$ is the worst-case coherence of $A$ and $[L]$ is the number of bits in a prefix code for $L$.

The proof of this theorem is provided in Appendix A.3.1. Note that the bound in (4.7) depends on (a) the signal-dependent interference term in (4.1) through $\|f\|_1$, (b) the variability of noise through $|c_u - c_\ell|$, (c) the choice of $A$ through $\mu(A)$, (d) the depth of each leaf through $[L]$, (e) the size of each leaf through $\sum_{i,j \in L} \langle A^{(i)}, A^{(j)} \rangle$ and (f) the parameter $\delta$. As a result, searching for an estimate $\hat{S} \in S$ that minimizes the upper bound of $R(S)$ in (4.7) will favor estimates with few, deep leaves that hone in on the boundary of the level set. Though the theoretical analysis of our method is significantly different from the analysis in [44] because of the interference term, it only changes the way the penalty is defined in our setup. As a result, we can adapt the computational techniques discussed in [44] to compute our estimator in an efficient way. Our method is computationally efficient since the proxy computation needs at most $O(KN)$ operations (fewer if $A$ contains certain structure; e.g., $A$ is a Toeplitz matrix) and the level set estimation method needs $O(N \log N)$ operations.
4.3.1 Performance improvement via DC offset subtraction

So far we showed that the signal-dependent interference term in (4.1) leads to a penalty term proportional to \( \|f\|_1 \) in (4.7). This implies that the interference in \( z \) and thus the performance of our method worsens as the signal energy of \( f \) increases. To find a way to minimize the signal-dependent interference, let us write \( f = \tilde{f} + \lambda \mathbb{1} \), where \( \lambda \) is a constant DC offset such that

\[
\|\tilde{f}\|_1 \leq \|f\|_1.
\]

(4.8)

If we have access to an estimate \( \hat{\lambda} \) of \( \lambda \), then we can minimize the signal-dependent interference by subtracting this DC offset to obtain

\[
\tilde{y} = y - A\hat{\lambda} = A(\tilde{f} + \lambda \mathbb{1}) + n - A\hat{\lambda} = A\left(\tilde{f} + (\lambda - \hat{\lambda}) \mathbb{1}\right) + n \approx A\hat{f} + n,
\]

assuming that \( \hat{\lambda} \approx \lambda \). The proxy observations in this case reduce to

\[
\tilde{z} = A^T\tilde{y} \approx A^TA\hat{f} + A^TAn = \tilde{f} + (A^TA - I)\hat{f} + A^TAn.
\]

Since \( S^* = \{i : f_i > \gamma\} \equiv \{i : \tilde{f}_i > \tilde{\gamma}\} \), where \( \tilde{\gamma} = \gamma - \lambda \), we can estimate \( S^* \) from \( \tilde{z} \) using our level set estimation method discussed in this section.

If we let \( \lambda \) to be the median of \( f \), then we can easily see that (4.8) holds for this particular choice of \( \lambda \). In practice, however, estimation of the median of \( f \) from \( y \) might be very hard, though the estimation of the mean of \( f \) might be tractable. For instance, if we construct \( A' = [\mathbb{1}^T] \) such that the first row of \( A' \) is \( \mathbb{1}^T \), then \( y' = A'f + n = [y_1'] \), and \( \hat{\lambda} = y_1'/N = (\sum_i f_i + n_1)/N = \lambda + n_1/N \). If the sensor noise is negligible, or if \( N \) is large, then \( \hat{\lambda} \approx \lambda \) and we can perform mean subtraction, instead of median subtraction to reduce the signal-dependent interference. While (4.8) does not always hold if \( \lambda \) is the mean of \( f \), simulation results in Sec. 4.5 suggest that mean subtraction can result in significant improvement in performance.
Recent work on sparse support estimation in [41] suggests that the second term (proportional to $\mu(A)\|f\|_1$) in the right hand side of (4.7) may be tightened significantly if we make some additional assumptions. In particular, if $f$ is assumed to be approximately sparse, so that the $\ell_1$ norm of $f$ restricted to the set $\mathcal{S}^*$ is sufficiently small, then the bound on $T_1$ may be tightened by using the average coherence

$$\nu(A) = \frac{1}{N-1}\|A^TA - I\|_\infty$$

of the projection operator $A$ [41]. A comprehensive analysis that utilizes both $\mu(A)$ and $\nu(A)$ to tighten the bounds in Theorem 1 will be explored in a sequel to this work.

4.4 Relationship with wavelet-based methods

The success of wavelet-based methods in estimating a piecewise smooth function from noisy measurements suggests a potential extension of such methods to the problem of level set estimation [55]. For instance, one possible approach for level set estimation from projection measurements is to first estimate the underlying function $f$ from proxy measurements $z$ using wavelet-based denoising methods and then threshold the resulting estimate at level $\gamma$. Estimating $f$ from $y$ through an intermediate proxy construction step is similar to the iterative hard thresholding method in compressive sensing literature with just one iteration [108]. While such ‘plug-in’ estimation techniques using wavelet-based methods offer practical solutions to the level set estimation problem, their estimation performances are very hard to characterize. The iterative hard thresholding methods aim to minimize the mean squared error over the entire image. This, however, does not attempt to minimize the errors close to the level set boundaries, which is critical to the characterization of level set estimation performance.

The proposed multiscale, partition based set estimation method with proxy measurements can be thought of as a combination of the iterative hard thresholding method with just one iteration, and wavelet-based denoising ideas. Specifically, our
partition-based method is similar in spirit to the wavelet-based denoising ideas using the unnormalized Haar wavelet transform. Both wavelet-based methods and our method rely on the spatial homogeneity of the underlying function $f$ to perform level set estimation. The difference between the two methods stems from the way in which the wavelet coefficients are thresholded in each case. While the wavelet-based method thresholds are chosen to minimize the mean squared error, our method thresholds the coefficients at levels that are tailored to the level set estimation problem. Since the proposed method shares similar ideas with wavelet-based methods, the proof techniques presented in this chapter could potentially be extended to wavelet-based methods in order to characterize their estimation performances.

Due to the lack of a theoretical performance comparison between the wavelet-based methods and our method, we present an empirical comparison of these methods in the following section by conducting experiments on several test images. Preliminary experimental results discussed below demonstrate that the proposed partition-based, multiscale method using proxy observations (a) is a powerful tool to perform direct level set estimation from projection measurements, (b) allows us to exploit the spatial homogeneity of the underlying function to perform set estimation, (c) performs an order of magnitude better than thresholding methods that obtain level set estimates by simply thresholding the proxy observations at level $\gamma$, and (d) yields results that are comparable to the results obtained using wavelet-based thresholding approaches for most test images considered in these experiments.

4.5 Experimental results

In order to test the effectiveness of our level set estimation algorithm, we conduct experiments on fifteen different images; eight of which correspond to magnetic resonance images obtained from the publicly available PubMed image database (http://www.ncbi.nlm.nih.gov/pmc/articles/
five of the test images are obtained from the publicly available NASA’s Wide-field Infrared Survey Explorer (WISE) webpage (http://wise.ssl.berkeley.edu/gallery_images.html), and two of the test images correspond to functions of the Hyperspectral Mapper (HyMap) data. To facilitate a transparent discussion, we refer to the medical images by BioImage1 through BioImage8; the astronomical images by AstroImage1 through AstroImage5 and the HyMap images by HyMapImage1 and HyMapImage2 respectively. Each test image is of size $128 \times 128$. In these experiments, we are interested in estimating $\gamma$-level set of each test image from noisy, projection measurements of the form $\mathbf{y} = \mathbf{Af} + \mathbf{n} \in \mathbb{R}^K$ for $K < N = 128 \times 128$, without reconstructing $\mathbf{f}$ from $\mathbf{y}$. The entries of the projection operator in these experiments are drawn from $\sim \mathcal{N}(0,1/K)$ and $\mathbf{n} \sim \mathcal{N}(0,\mathbf{I})$. Note that we consider a Gaussian noise model here, which is unbounded. However, it is bounded with high probability and our theory can also be extended to this case.

We compare the performance of our method with the performances of the following approaches using the excess risk error metric defined in (4.2): 

(a) **Thresholding method**, where the estimate $\hat{S}_\gamma$ is simply obtained by thresholding the proxy observations $\mathbf{z}$ at level $\gamma$, that is, $\hat{S}_\gamma = \{i : z_i \geq \gamma\}$,

(b) **Risk-optimal thresholding method**, where the estimate $\hat{S}_{\hat{\gamma}}$ is obtained by thresholding $\mathbf{z}$ at a level $\hat{\gamma}$ that minimizes the excess risk, that is, $\hat{S}_{\hat{\gamma}} = \{i : z_i \geq \hat{\gamma}\}$ where $\hat{\gamma} = \arg \min_{\gamma} \mathcal{E}(\hat{S}_\gamma, S^*)$, and

(c) **Wavelet-based method**, where the estimate $\hat{S}_w$ is obtained by first estimating $\mathbf{f}$
from \( z \) using translation invariant (TI) wavelet denoising, and then thresholding the resulting estimate \( \hat{f} \) at level \( \gamma \), that is, \( \hat{S}_w = \{ i : \hat{f}_i \geq \gamma \} \). In these experiments we perform wavelet denoising using Daubechies-4 wavelets and soft thresholding, where the threshold is chosen to minimize the excess risk.

In these simulation experiments, we compute the excess risk clairvoyantly based on the knowledge of \( f \). We obtain the estimates \( \hat{S} \) using our method by minimizing (4.7) with a scaling factor in front of the sum over leafs, which is chosen to minimize \( \bar{\xi}(\hat{s}, s^*) \).

Each image, and the corresponding level set threshold, in our set of test images are chosen to study the performances of competing methods under different challenges. For instance, some of the test images have level sets that are smooth and diffuse (e.g., BioImage5, AstroImage3), some have level sets with irregular boundaries (e.g., BioImage3, BioImage7, BioImage8, AstroImage4, AstroImage5, HyMapImage1 and HyMapImage2), and some other have level sets that have a combination of smooth and irregular boundaries (e.g., BioImage1, BioImage2, BioImage4, BioImage6, AstroImage1, AstroImage2). Figs. 4.1(a), 4.3(a), 4.5(a), 4.7(a), 4.9(a), 4.11(a), 4.13(a), 4.15(a), 4.17(a), 4.19(a), 4.21(a), 4.23(a), 4.25(a), 4.27(a), and 4.29(a) show the test images considered in this experiment. Figs. 4.1(b), 4.3(b), 4.5(b), 4.7(b), 4.9(b), 4.11(b), 4.13(b), 4.15(b), 4.17(b), 4.19(b), 4.21(b), 4.23(b), 4.25(b), 4.27(b), and 4.29(b) show the level sets corresponding to each test image. For each test case, we obtained plots of excess risk versus different values of \( K < N \) by averaging the results obtained for 100 different noise and projection matrix realizations, with and without performing mean subtraction. Figs. 4.1(c), 4.3(c), 4.5(c), 4.7(c), 4.9(c), 4.11(c), 4.13(c), 4.15(c), 4.17(c), 4.19(c), 4.21(c), 4.23(c), 4.25(c), 4.27(c), and 4.29(c) show the plots of excess risk versus \( K \) obtained for each test image without performing mean subtraction. Figs. 4.1(d), 4.3(d), 4.5(d), 4.7(d), 4.9(d), 4.11(d), 4.13(d),
4.15(d), 4.17(d), 4.19(d), 4.21(d), 4.23(d), 4.25(d), 4.27(d), and 4.29(d) show the plots of excess risk versus $K$ obtained for each test image after performing mean subtraction. These results suggest that mean subtraction results in a significant performance improvement in each test case for all four competing methods considered in these experiments. This is largely due to the fact that the mean subtraction operation minimizes the signal-dependent interference term, which is a by-product of our proxy construction. This is also evident by comparing snapshots of the proxy measurements of each test image without performing mean subtraction and after performing mean subtraction when $K = N/2$, as shown in Figs. 4.2(a-b), 4.4(a-b), 4.6(a-b), 4.8(a-b), 4.10(a-b), 4.12(a-b), 4.14(a-b), 4.16(a-b), 4.18(a-b), 4.20(a-b), 4.22(a-b), 4.24(a-b), 4.26(a-b), 4.28(a-b), 4.30(a-b) respectively.

Note that each entry of the proxy measurements can be written as $z_i = f_i + n'_i$, where $n'_i = \sum_{j \neq i} f_j \langle A^{(i)}, A^{(j)} \rangle + \langle A^{(i)}, n \rangle$. As a result, thresholding $z$ at level $\gamma$ will result in several false positives and misses at locations where $n'_i$ is comparable to $f_i$, as shown in Figs. 4.2(c-d), 4.4(c-d), 4.6(c-d), 4.8(c-d), 4.10(c-d), 4.12(c-d), 4.14(c-d), 4.16(c-d), 4.18(c-d), 4.20(c-d), 4.22(c-d), 4.24(c-d), 4.26(c-d), 4.28(c-d), 4.30(c-d) respectively. The risk-optimal thresholding method seems to minimize the number of errors that the thresholding method yields, but still fails to estimate the level sets successfully as shown in Figs. 4.2(e-f), 4.4(e-f), 4.6(e-f), 4.8(e-f), 4.10(e-f), 4.12(e-f), 4.14(e-f), 4.16(e-f), 4.18(e-f), 4.20(e-f), 4.22(e-f), 4.24(e-f), 4.26(e-f), 4.28(e-f), 4.30(e-f) respectively. Compared to the results obtained using the thresholding method and the risk-optimal thresholding method, the wavelet-based method yields better results because of the spatial adaptivity of wavelet-based estimation methods. Figs. 4.2(g-h), 4.4(g-h), 4.6(g-h), 4.8(g-h), 4.10(g-h), 4.12(g-h), 4.14(g-h), 4.16(g-h), 4.18(g-h), 4.20(g-h), 4.22(g-h), 4.24(g-h), 4.26(g-h), 4.28(g-h), 4.30(g-h) show snapshots of the results obtained using the wavelet-based thresholding method when $K = N/2$, without and after performing mean subtraction. Figs. 4.2(i-j), 4.4(i-
j), 4.6(i-j), 4.8(i-j), 4.10(i-j), 4.12(i-j), 4.14(i-j), 4.16(i-j), 4.18(i-j), 4.20(i-j), 4.22(i-j),
4.24(i-j), 4.26(i-j), 4.28(i-j), 4.30(i-j) show snapshots of the results obtained for each
test case using our method when $K = N/2$. These preliminary results, together
with the plots of excess risk versus $K$ in Figs. 4.1(c-d), 4.3(c-d), 4.5(c-d), 4.7(c-
d), 4.9(c-d), 4.11(c-d), 4.13(c-d), 4.15(c-d), 4.17(c-d), 4.19(c-d), 4.21(c-d), 4.23(c-d),
4.25(c-d), 4.27(c-d), and 4.29(c-d) suggest that our method yields results that are
comparable to those obtained using wavelet-based thresholding approaches for all
test images considered in our experiments and in many cases, slightly better than
the wavelet-based methods, especially when $K \ll N$. Since our method exploits the
spatial piecewise homogeneity of the underlying function $f$, it effectively minimizes
the number of false positives and misses that are seen in the results obtained using
competing methods. This is also the reason why our method misses isolated pixels
corresponding to some level sets level sets (e.g., results corresponding to BioImage3,
BioImage4, BioImage6, BioImage7, BioImage8, AstroImage1, AstroImage2,
AstroImage4). Note also that the wavelet-based method also has similar difficulties
in estimating isolated level set pixels from proxy observations since neighboring pix-
 els cannot be successfully exploited in such a setting to perform spatially-adaptive
level set estimation.

It is insightful to compare the performances of the wavelet-based method and
our method on different test images that correspond to level sets with different chal-
lenges. The wavelet-based method seems to outperform our method if the level
sets are smooth and diffuse as shown in the results corresponding to AstroImage3
(see Figs. 4.21(c-d)). This is unsurprising since wavelet-based estimation methods
are shown to be effective in estimating images that have smooth surfaces separated
by smooth boundaries. However, when the level set boundaries are more irregular,
(e.g., BioImage3, BioImage7, AstroImage4, HyMapImage1, HyMapImage2), then
our partition-based level set estimation method better adapts to the level set bound-
aries since our method encourages estimates with few, deep leaves that hone in on the level set boundary.

4.6 Conclusion

This work proposes a theoretically tractable and computationally efficient tree-based approach to extract level sets of a function from projection measurements without reconstructing the underlying function. The preliminary results presented in Sec. 4.5 suggest that the proposed methods yield promising results in several applications such as medical imaging, astronomical imaging and spectral imaging. One of the key advantages of our approach is that we can parallelize the level set estimation problem when the domain of the function of interest is very large. In such cases, we can partition the data into different patches, run our estimation algorithm on each patch separately and merge the results to identify the regions that correspond to the level set. In applications such as medical imaging, the time saved by collecting fewer projective measurements and parallelization can be significant and crucial.

The methods and analysis presented in this chapter yield a first pass at the problem of level set estimation from projection measurements without an intermediate reconstruction step. Since our methods are similar in spirit with the wavelet-based methods, our proof techniques could potentially lead to a more thorough performance analysis of our method, as well as the wavelet-based level set estimation methods.
Figure 4.1: (a) Level set experimental results corresponding to BioImage1. (a) True function $f \in \mathbb{R}^{128 \times 128}$ such that $f_i \in [87, 765]$. We measure $K$ Gaussian random projections of this image. (b) Level set $S^* = \{i : f_i \geq 420\}$ such that $|S^*| \approx 0.088N$ where $N = 128 \times 128$. (c) Plot of excess risk as a function of $K < N = 16384$ without performing mean subtraction. (d) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.
### Results obtained with $K = N/2$ measurements

<table>
<thead>
<tr>
<th></th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy observations $\tilde{z}$</td>
<td><img src="a" alt="Figure" /></td>
<td><img src="b" alt="Figure" /></td>
</tr>
<tr>
<td>Estimates obtained by the thresholding method</td>
<td>$\varepsilon(\hat{S}_\gamma, S^*) = 58.63$</td>
<td>$\varepsilon(\hat{S}_\gamma, S^*) = 26.97$</td>
</tr>
<tr>
<td></td>
<td><img src="c" alt="Figure" /></td>
<td><img src="d" alt="Figure" /></td>
</tr>
<tr>
<td>Estimates obtained by the risk-optimal thresholding method</td>
<td>$\varepsilon(\hat{S}_\gamma, S^*) = 18.71$</td>
<td>$\varepsilon(\hat{S}_\gamma, S^*) = 10.13$</td>
</tr>
<tr>
<td></td>
<td><img src="e" alt="Figure" /></td>
<td><img src="f" alt="Figure" /></td>
</tr>
<tr>
<td>Estimates obtained by the wavelet-based method</td>
<td>$\varepsilon(\hat{S}_w, S^*) = 4.792$</td>
<td>$\varepsilon(\hat{S}_w, S^*) = 1.675$</td>
</tr>
<tr>
<td></td>
<td><img src="g" alt="Figure" /></td>
<td><img src="h" alt="Figure" /></td>
</tr>
<tr>
<td>Estimates obtained by our method</td>
<td>$\varepsilon(\hat{s}, S^*) = 3.955$</td>
<td>$\varepsilon(\hat{s}, S^*) = 1.544$</td>
</tr>
<tr>
<td></td>
<td><img src="i" alt="Figure" /></td>
<td><img src="j" alt="Figure" /></td>
</tr>
</tbody>
</table>

**Figure 4.2:** Snapshots of level set estimation simulation results corresponding to BioImage1 when $K = N/2$. 

90
Figure 4.3: (a) Level set experimental results corresponding to BioImage2. (a) True function $f \in \mathbb{R}^{128 \times 128}$ such that $f_i \in [18, 255]$. We measure $K$ Gaussian random projections of this image. (b) Level set $S^* = \{ i : f_i \geq 155 \}$ such that $|S^*| \approx 0.1772N$ where $N = 128 \times 128$. (c) Plot of excess risk as a function of $K < N = 16384$ without performing mean subtraction. (d) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.
### Results obtained with $K = N/2$ measurements

<table>
<thead>
<tr>
<th></th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy observations $\tilde{z}$</td>
<td><img src="image_a" alt="Proxy observations without mean subtraction" /></td>
<td><img src="image_b" alt="Proxy observations with mean subtraction" /></td>
</tr>
<tr>
<td>Estimates obtained by the thresholding method</td>
<td><img src="image_c" alt="Thresholding estimates without mean subtraction" /></td>
<td>$\varepsilon(\hat{S}, S^*) = 23.87$</td>
</tr>
<tr>
<td>Estimates obtained by the risk-optimal thresholding method</td>
<td><img src="image_e" alt="Risk-optimal thresholding estimates without mean subtraction" /></td>
<td>$\varepsilon(\hat{S}, S^*) = 11.66$</td>
</tr>
<tr>
<td>Estimates obtained by the wavelet-based method</td>
<td><img src="image_g" alt="Wavelet-based estimates without mean subtraction" /></td>
<td>$\varepsilon(\hat{S}, S^*) = 1.861$</td>
</tr>
<tr>
<td>Estimates obtained by our method</td>
<td><img src="image_i" alt="Our method estimates without mean subtraction" /></td>
<td>$\varepsilon(\hat{S}, S^*) = 1.490$</td>
</tr>
</tbody>
</table>

**Figure 4.4**: Snapshots of level set estimation simulation results corresponding to BioImage2 when $K = N/2$. 

92
Figure 4.5: (a) Level set experimental results corresponding to BioImage3. (a) True function $f \in \mathbb{R}^{128 \times 128}$ such that $f_i \in [16, 254]$. We measure $K$ Gaussian random projections of this image. (b) Level set $S^* = \{i : f_i \geq 155\}$ such that $|S^*| \approx 0.2294N$ where $N = 128 \times 128$. (c) Plot of excess risk as a function of $K < N = 16384$ without performing mean subtraction. (d) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.
Results obtained with $K = N/2$ measurements

<table>
<thead>
<tr>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy observations $\tilde{z}$</td>
<td><img src="a" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the thresholding method</td>
<td><img src="c" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the risk-optimal thresholding method</td>
<td><img src="e" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the wavelet-based method</td>
<td><img src="g" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by our method</td>
<td><img src="i" alt="Image" /></td>
</tr>
</tbody>
</table>

**Figure 4.6**: Snapshots of level set estimation simulation results corresponding to BioImage3 when $K = N/2$. 

94
Figure 4.7: (a) Level set experimental results corresponding to BioImage4. (a) True function $f \in \mathbb{R}^{128 \times 128}$ such that $f_i \in [17, 254]$. We measure $K$ Gaussian random projections of this image. (b) Level set $S^* = \{i : f_i \geq 170\}$ such that $|S^*| \approx 0.1169N$ where $N = 128 \times 128$. (c) Plot of excess risk as a function of $K < N = 16384$ without performing mean subtraction. (d) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.
Results obtained with $K = N/2$ measurements

<table>
<thead>
<tr>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy observations $\tilde{z}$</td>
<td><img src="a" alt="Image" /> <img src="b" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the thresholding method</td>
<td><img src="c" alt="Image" /> <img src="d" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_{\gamma},S^*) = 21.63$</td>
<td>$\varepsilon(\hat{S}_{\gamma},S^*) = 10.68$</td>
</tr>
<tr>
<td>Estimates obtained by the risk-optimal thresholding method</td>
<td><img src="e" alt="Image" /> <img src="f" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_{\gamma},S^*) = 6.142$</td>
<td>$\varepsilon(\hat{S}_{\gamma},S^*) = 4.990$</td>
</tr>
<tr>
<td>Estimates obtained by the wavelet-based method</td>
<td><img src="g" alt="Image" /> <img src="h" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_{w},S^*) = 1.827$</td>
<td>$\varepsilon(\hat{S}_{w},S^*) = 0.9067$</td>
</tr>
<tr>
<td>Estimates obtained by our method</td>
<td><img src="i" alt="Image" /> <img src="j" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S},S^*) = 1.653$</td>
<td>$\varepsilon(\hat{S},S^*) = 0.8529$</td>
</tr>
</tbody>
</table>

Figure 4.8: Snapshots of level set estimation simulation results corresponding to BioImage4 when $K = N/2$. 

96
Figure 4.9: (a) Level set experimental results corresponding to BioImage5. (a) True function $\mathbf{f} \in \mathbb{R}^{128 \times 128}$ such that $f_i \in [0, 255]$. We measure $K$ Gaussian random projections of this image. (b) Level set $S^* = \{i : f_i \geq 150\}$ such that $|S^*| \approx 0.0425N$ where $N = 128 \times 128$. (c) Plot of excess risk as a function of $K < N = 16384$ without performing mean subtraction. (d) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.
Results obtained with $K = N/2$ measurements

<table>
<thead>
<tr>
<th></th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proxy observations $\tilde{z}$</strong></td>
<td><img src="a" alt="Image" /></td>
<td><img src="b" alt="Image" /></td>
</tr>
<tr>
<td><strong>Estimates obtained by the thresholding method</strong></td>
<td><img src="c" alt="Image" /></td>
<td><img src="d" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_{\gamma}, S^*) = 12.41$</td>
<td>$\varepsilon(\hat{S}_{\gamma}, S^*) = 6.649$</td>
<td></td>
</tr>
<tr>
<td><strong>Estimates obtained by the risk-optimal thresholding method</strong></td>
<td><img src="e" alt="Image" /></td>
<td><img src="f" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_{\gamma}, S^*) = 1.761$</td>
<td>$\varepsilon(\hat{S}_{\gamma}, S^*) = 1.597$</td>
<td></td>
</tr>
<tr>
<td><strong>Estimates obtained by the wavelet-based method</strong></td>
<td><img src="g" alt="Image" /></td>
<td><img src="h" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_w, S^*) = 0.2371$</td>
<td>$\varepsilon(\hat{S}_w, S^*) = 0.1342$</td>
<td></td>
</tr>
<tr>
<td><strong>Estimates obtained by our method</strong></td>
<td><img src="i" alt="Image" /></td>
<td><img src="j" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}, S^*) = 0.1354$</td>
<td>$\varepsilon(\hat{S}, S^*) = 0.1068$</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.10:** Snapshots of level set estimation simulation results corresponding to BioImage5 when $K = N/2$. 

98
Figure 4.11: (a) Level set experimental results corresponding to BioImage6. (a) True function $f \in \mathbb{R}^{128 \times 128}$ such that $f_i \in [0, 255]$. We measure $K$ Gaussian random projections of this image. (b) Level set $S^* = \{i : f_i \geq 130\}$ such that $|S^*| \approx 0.0821N$ where $N = 128 \times 128$. (c) Plot of excess risk as a function of $K < N = 16384$ without performing mean subtraction. (d) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.
Results obtained with $K = N/2$ measurements

<table>
<thead>
<tr>
<th></th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy observations $\tilde{z}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a)</td>
<td></td>
<td>(b)</td>
</tr>
<tr>
<td>Estimates obtained by the thresholding method</td>
<td>$\varepsilon(\hat{S}_\gamma, S^*) = 14.65$</td>
<td>$\varepsilon(\hat{S}_\gamma, S^*) = 8.551$</td>
</tr>
<tr>
<td>(c)</td>
<td></td>
<td>(d)</td>
</tr>
<tr>
<td>Estimates obtained by the risk-optimal thresholding method</td>
<td>$\varepsilon(\hat{S}_\hat{\gamma}, S^*) = 2.742$</td>
<td>$\varepsilon(\hat{S}_\hat{\gamma}, S^*) = 2.483$</td>
</tr>
<tr>
<td>(e)</td>
<td></td>
<td>(f)</td>
</tr>
<tr>
<td>Estimates obtained by the wavelet-based method</td>
<td>$\varepsilon(\hat{S}_w, S^*) = 0.9086$</td>
<td>$\varepsilon(\hat{S}_w, S^*) = 0.7260$</td>
</tr>
<tr>
<td>(g)</td>
<td></td>
<td>(h)</td>
</tr>
<tr>
<td>Estimates obtained by our method</td>
<td>$\varepsilon(\hat{S}, S^*) = 0.7513$</td>
<td>$\varepsilon(\hat{S}, S^*) = 0.7255$</td>
</tr>
<tr>
<td>(i)</td>
<td></td>
<td>(j)</td>
</tr>
</tbody>
</table>

**Figure 4.12**: Snapshots of level set estimation simulation results corresponding to BioImage6 when $K = N/2$. 

100
Our method
Wavelet-based method
Thresholding
Customized thresholding
Excess risk

Figure 4.13: (a) Level set experimental results corresponding to BioImage7. (a) True function \( f \in \mathbb{R}^{128 \times 128} \) such that \( f_i \in [12, 255] \). We measure \( K \) Gaussian random projections of this image. (b) Level set \( S^* = \{ i : f_i \geq 130 \} \) such that \( |S^*| \approx 0.1710N \) where \( N = 128 \times 128 \). (c) Plot of excess risk as a function of \( K < N = 16384 \) without performing mean subtraction. (d) Plot of excess risk as a function of \( K < N = 16384 \) after performing mean subtraction.
Results obtained with $K = N/2$ measurements

<table>
<thead>
<tr>
<th></th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy observations $\tilde{z}$</td>
<td><img src="a" alt="Image" /></td>
<td><img src="b" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the thresholding method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_\gamma, S^*) = 15.66$</td>
<td><img src="c" alt="Image" /></td>
<td><img src="d" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_\gamma, S^*) = 8.708$</td>
<td><img src="e" alt="Image" /></td>
<td><img src="f" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the risk-optimal thresholding method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_\hat{\gamma}, S^*) = 7.794$</td>
<td><img src="g" alt="Image" /></td>
<td><img src="h" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_\hat{\gamma}, S^*) = 4.868$</td>
<td><img src="i" alt="Image" /></td>
<td><img src="j" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the wavelet-based method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_w, S^*) = 2.756$</td>
<td><img src="k" alt="Image" /></td>
<td><img src="l" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_w, S^*) = 1.174$</td>
<td><img src="m" alt="Image" /></td>
<td><img src="n" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by our method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\varepsilon(\hat{s}, S^*) = 2.291$</td>
<td><img src="o" alt="Image" /></td>
<td><img src="p" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{s}, S^*) = 1.119$</td>
<td><img src="q" alt="Image" /></td>
<td><img src="r" alt="Image" /></td>
</tr>
</tbody>
</table>

**Figure 4.14:** Snapshots of level set estimation simulation results corresponding to BioImage7 when $K = N/2$.  

102
Figure 4.15: (a) Level set experimental results corresponding to BioImage8. (a) True function \( f \in \mathbb{R}^{128 \times 128} \) such that \( f_i \in [4, 255] \). We measure \( K \) Gaussian random projections of this image. (b) Level set \( S^* = \{i : f_i \geq 140\} \) such that \( |S^*| \approx 0.1127N \) where \( N = 128 \times 128 \). (c) Plot of excess risk as a function of \( K < N = 16384 \) without performing mean subtraction. (d) Plot of excess risk as a function of \( K < N = 16384 \) after performing mean subtraction.
Results obtained with $K = N/2$ measurements

<table>
<thead>
<tr>
<th>Proxy observations $\tilde{z}$</th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="a" alt="Image" /></td>
<td><img src="b" alt="Image" /></td>
<td></td>
</tr>
</tbody>
</table>

Estimates obtained by the thresholding method

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon(\hat{S}_{\gamma},S^*)$</td>
<td>19.93</td>
<td>9.992</td>
</tr>
<tr>
<td><img src="c" alt="Image" /></td>
<td><img src="d" alt="Image" /></td>
<td></td>
</tr>
</tbody>
</table>

Estimates obtained by the risk-optimal thresholding method

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon(\hat{S}_{\gamma},S^*)$</td>
<td>5.6259</td>
<td>4.026</td>
</tr>
<tr>
<td><img src="e" alt="Image" /></td>
<td><img src="f" alt="Image" /></td>
<td></td>
</tr>
</tbody>
</table>

Estimates obtained by the wavelet-based method

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon(\hat{S}_{w},S^*)$</td>
<td>1.992</td>
<td>1.094</td>
</tr>
<tr>
<td><img src="g" alt="Image" /></td>
<td><img src="h" alt="Image" /></td>
<td></td>
</tr>
</tbody>
</table>

Estimates obtained by our method

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon(\hat{S},S^*)$</td>
<td>1.797</td>
<td>1.075</td>
</tr>
<tr>
<td><img src="i" alt="Image" /></td>
<td><img src="j" alt="Image" /></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.16: Snapshots of level set estimation simulation results corresponding to BioImage8 when $K = N/2$. 

104
Figure 4.17: (a) Level set experimental results corresponding to AstroImage1. (a) True function $f \in \mathbb{R}^{128 \times 128}$ such that $f_i \in [3, 760]$. We measure $K$ Gaussian random projections of this image. (b) Level set $S^* = \{i : f_i \geq 395\}$ such that $|S^*| \approx 0.1405N$ where $N = 128 \times 128$. (c) Plot of excess risk as a function of $K < N = 16384$ without performing mean subtraction. (d) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.
<table>
<thead>
<tr>
<th></th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy observations ( \tilde{z} )</td>
<td><img src="a" alt="Image" /></td>
<td><img src="b" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the thresholding method</td>
<td><img src="c" alt="Image" />  ( \varepsilon(\hat{S}_\gamma, S^*) = 48.0847 )</td>
<td><img src="d" alt="Image" />  ( \varepsilon(\hat{S}_\gamma, S^*) = 21.5347 )</td>
</tr>
<tr>
<td>Estimates obtained by the risk-optimal thresholding method</td>
<td><img src="e" alt="Image" />  ( \varepsilon(\hat{S}_\gamma, S^*) = 8.258 )</td>
<td><img src="f" alt="Image" />  ( \varepsilon(\hat{S}_\gamma, S^*) = 7.194 )</td>
</tr>
<tr>
<td>Estimates obtained by the wavelet-based method</td>
<td><img src="g" alt="Image" />  ( \varepsilon(\hat{S}_w, S^*) = 3.254 )</td>
<td><img src="h" alt="Image" />  ( \varepsilon(\hat{S}_w, S^*) = 2.245 )</td>
</tr>
<tr>
<td>Estimates obtained by our method</td>
<td><img src="i" alt="Image" />  ( \varepsilon(\hat{S}, S^*) = 1.489 )</td>
<td><img src="j" alt="Image" />  ( \varepsilon(\hat{S}, S^*) = 0.9887 )</td>
</tr>
</tbody>
</table>

**Figure 4.18:** Snapshots of level set estimation simulation results corresponding to AstroImage1 when \( K = N/2 \).
Figure 4.19: (a) Level set experimental results corresponding to AstroImage2. (a) True function $f \in \mathbb{R}^{128 \times 128}$ such that $f_i \in [63, 765]$. We measure $K$ Gaussian random projections of this image. (b) Level set $S^* = \{i : f_i \geq 580\}$ such that $|S^*| \approx 0.1284N$ where $N = 128 \times 128$. (c) Plot of excess risk as a function of $K < N = 16384$ without performing mean subtraction. (d) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.
Results obtained with $K = N/2$ measurements

<table>
<thead>
<tr>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
</table>

Proxy observations $\tilde{z}$

(a) ![Proxy observations without mean subtraction](image)

(b) ![Proxy observations with mean subtraction](image)

Estimates obtained by the thresholding method

$c) \hat{S}_{\gamma}^{*}, S^*$

$\epsilon(\hat{S}_{\gamma}, S^*) = 65.1887$

(d) \hat{S}_{\gamma}^{*}, S^*$

$\epsilon(\hat{S}_{\gamma}, S^*) = 26.85$

Estimates obtained by the risk-optimal thresholding method

(e) \hat{S}_{\gamma}^{*}, S^*$

$\epsilon(\hat{S}_{\gamma}, S^*) = 8.1703$

(f) \hat{S}_{\gamma}^{*}, S^*$

$\epsilon(\hat{S}_{\gamma}, S^*) = 7.9349$

Estimates obtained by the wavelet-based method

(g) \hat{S}_{w}^{*}, S^*$

$\epsilon(\hat{S}_{w}, S^*) = 3.403$

(h) \hat{S}_{w}^{*}, S^*$

$\epsilon(\hat{S}_{w}, S^*) = 1.319$

Estimates obtained by our method

(i) $\hat{S}, S^*$

$\epsilon(\hat{S}, S^*) = 2.644$

(j) $\hat{S}, S^*$

$\epsilon(\hat{S}, S^*) = 1.337$

**Figure 4.20:** Snapshots of level set estimation simulation results corresponding to AstroImage2 when $K = N/2$. 

108
Figure 4.21: (a) Level set experimental results corresponding to AstroImage3. 
(a) True function \( f \in \mathbb{R}^{128 \times 128} \) such that \( f_i \in [41, 765] \). We measure \( K \) Gaussian random projections of this image. (b) Level set \( S^* = \{i : f_i \geq 380\} \) such that \( |S^*| \approx 0.06090N \) where \( N = 128 \times 128 \). (c) Plot of excess risk as a function of \( K < N = 16384 \) without performing mean subtraction. (d) Plot of excess risk as a function of \( K < N = 16384 \) after performing mean subtraction.
Results obtained with $K = N/2$ measurements

<table>
<thead>
<tr>
<th></th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy observations $\tilde{z}$</td>
<td><img src="a" alt="Image" /></td>
<td><img src="b" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the thresholding method</td>
<td><img src="c" alt="Image" /></td>
<td><img src="d" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td>$\varepsilon(\hat{S}_{\gamma}, S^*) = 41.94$</td>
<td>$\varepsilon(\hat{S}_{\gamma}, S^*) = 12.51$</td>
</tr>
<tr>
<td>Estimates obtained by the risk-optimal thresholding method</td>
<td><img src="e" alt="Image" /></td>
<td><img src="f" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td>$\varepsilon(\hat{S}_{\gamma}, S^*) = 4.579$</td>
<td>$\varepsilon(\hat{S}_{\gamma}, S^*) = 3.254$</td>
</tr>
<tr>
<td>Estimates obtained by the wavelet-based method</td>
<td><img src="g" alt="Image" /></td>
<td><img src="h" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td>$\varepsilon(\hat{S}_{w}, S^*) = 0.2775$</td>
<td>$\varepsilon(\hat{S}_{w}, S^*) = 0.2335$</td>
</tr>
<tr>
<td>Estimates obtained by our method</td>
<td><img src="i" alt="Image" /></td>
<td><img src="j" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td>$\varepsilon(\hat{S}, S^*) = 0.3289$</td>
<td>$\varepsilon(\hat{S}, S^*) = 0.2213$</td>
</tr>
</tbody>
</table>

**Figure 4.22:** Snapshots of level set estimation simulation results corresponding to AstroImage3 when $K = N/2$.  

110
Figure 4.23: (a) Level set experimental results corresponding to AstroImage4. (a) True function $f \in \mathbb{R}^{128 \times 128}$ such that $f_i \in [32, 750]$. We measure $K$ Gaussian random projections of this image. (b) Level set $S^* = \{ i : f_i \geq 400 \}$ such that $|S^*| \approx 0.1082 N$ where $N = 128 \times 128$. (c) Plot of excess risk as a function of $K < N = 16384$ without performing mean subtraction. (d) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.
Results obtained with $K = N/2$ measurements

<table>
<thead>
<tr>
<th></th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy observations $\tilde{z}$</td>
<td><img src="a" alt="Image" /></td>
<td><img src="b" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the thresholding method</td>
<td><img src="c" alt="Image" /></td>
<td><img src="d" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_{\gamma}, S^*) = 50.65$</td>
<td>$\varepsilon(\hat{S}_{\gamma}, S^*) = 19.01$</td>
<td></td>
</tr>
<tr>
<td>Estimates obtained by the risk-optimal thresholding method</td>
<td><img src="e" alt="Image" /></td>
<td><img src="f" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_{\gamma}, S^*) = 8.905$</td>
<td>$\varepsilon(\hat{S}_{\gamma}, S^*) = 6.945$</td>
<td></td>
</tr>
<tr>
<td>Estimates obtained by the wavelet-based method</td>
<td><img src="g" alt="Image" /></td>
<td><img src="h" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_w, S^*) = 4.045$</td>
<td>$\varepsilon(\hat{S}_w, S^*) = 2.339$</td>
<td></td>
</tr>
<tr>
<td>Estimates obtained by our method</td>
<td><img src="i" alt="Image" /></td>
<td><img src="j" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}, S^*) = 3.326$</td>
<td>$\varepsilon(\hat{S}, S^*) = 2.192$</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.24:** Snapshots of level set estimation simulation results corresponding to AstroImage4 when $K = N/2$.  

112
Figure 4.25: (a) Level set experimental results corresponding to AstroImage5. (a) True function $f \in \mathbb{R}^{128 \times 128}$ such that $f_i \in [85, 721]$. We measure $K$ Gaussian random projections of this image. (b) Level set $S^* = \{i : f_i \geq 500\}$ such that $|S^*| \approx 0.1969N$ where $N = 128 \times 128$. (c) Plot of excess risk as a function of $K < N = 16384$ without performing mean subtraction. (d) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.
<table>
<thead>
<tr>
<th>Results obtained with $K = N/2$ measurements</th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy observations $\tilde{z}$</td>
<td><img src="a" alt="Image" /></td>
<td><img src="b" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the thresholding method</td>
<td><img src="c" alt="Image" /></td>
<td><img src="d" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_{\gamma},S^*) = 58.74$</td>
<td>$\varepsilon(\hat{S}_{\gamma},S^*) = 25.63$</td>
<td></td>
</tr>
<tr>
<td>Estimates obtained by the risk-optimal thresholding method</td>
<td><img src="e" alt="Image" /></td>
<td><img src="f" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_{\theta},S^*) = 14.45$</td>
<td>$\varepsilon(\hat{S}_{\theta},S^*) = 12.68$</td>
<td></td>
</tr>
<tr>
<td>Estimates obtained by the wavelet-based method</td>
<td><img src="g" alt="Image" /></td>
<td><img src="h" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_{w},S^*) = 4.025$</td>
<td>$\varepsilon(\hat{S}_{w},S^*) = 1.088$</td>
<td></td>
</tr>
<tr>
<td>Estimates obtained by our method</td>
<td><img src="i" alt="Image" /></td>
<td><img src="j" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S},S^*) = 2.323$</td>
<td>$\varepsilon(\hat{S},S^*) = 0.9721$</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.26:** Snapshots of level set estimation simulation results corresponding to AstroImage5 when $K = N/2$.  

114
Our method
Wavelet-based method
Thresholding
Customized thresholding
Excess risk

Figure 4.27: (a) Level set experimental results corresponding to HyMapImage1. (a) True function $f \in \mathbb{R}^{128 \times 128}$ such that $f_i \in [36.22, 97.70]$. We measure $K$ Gaussian random projections of this image. (b) Level set $S^* = \{i : f_i \geq 60\}$ such that $|S^*| \approx 0.0577N$ where $N = 128 \times 128$. (c) Plot of excess risk as a function of $K < N = 16384$ without performing mean subtraction. (d) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.
<table>
<thead>
<tr>
<th>Results obtained with $K = N/2$ measurements</th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy observations $\tilde{z}$</td>
<td><img src="a" alt="Image" /></td>
<td><img src="b" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the thresholding method</td>
<td>$\varepsilon(\hat{S}_{\gamma,s^*}) = 10.26$</td>
<td>$\varepsilon(\hat{S}_{\gamma,s^*}) = 4.366$</td>
</tr>
<tr>
<td>Estimates obtained by the risk-optimal thresholding method</td>
<td>$\varepsilon(\hat{S}_{\hat{\gamma},s^*}) = 3.759$</td>
<td>$\varepsilon(\hat{S}_{\hat{\gamma},s^*}) = 1.825$</td>
</tr>
<tr>
<td>Estimates obtained by the wavelet-based method</td>
<td>$\varepsilon(\hat{S}_{w,s^*}) = 2.334$</td>
<td>$\varepsilon(\hat{S}_{w,s^*}) = 0.6764$</td>
</tr>
<tr>
<td>Estimates obtained by our method</td>
<td>$\varepsilon(\hat{S},s^*) = 1.806$</td>
<td>$\varepsilon(\hat{S},s^*) = 0.5616$</td>
</tr>
</tbody>
</table>

Figure 4.28: Snapshots of level set estimation simulation results corresponding to HyMapImage1 when $K = N/2$.  

116
Figure 4.29: (a) Level set experimental results corresponding to HyMapImage2. (a) True function $f \in \mathbb{R}^{128 \times 128}$ such that $f_i \in [36.08, 120]$. We measure $K$ Gaussian random projections of this image. (b) Level set $S^* = \{i : f_i \geq 70\}$ such that $|S^*| \approx 0.1969N$ where $N = 128 \times 128$. (c) Plot of excess risk as a function of $K < N = 16384$ without performing mean subtraction. (d) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.
Results obtained with $K = N/2$ measurements

<table>
<thead>
<tr>
<th></th>
<th>Without mean subtraction</th>
<th>With mean subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proxy observations $\tilde{z}$</td>
<td><img src="a" alt="Image" /></td>
<td><img src="b" alt="Image" /></td>
</tr>
<tr>
<td>Estimates obtained by the thresholding method</td>
<td><img src="c" alt="Image" /></td>
<td><img src="d" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_\gamma, S^*) = 7.504$</td>
<td>$\varepsilon(\hat{S}_\gamma, S^*) = 2.616$</td>
<td></td>
</tr>
<tr>
<td>Estimates obtained by the risk-optimal thresholding method</td>
<td><img src="e" alt="Image" /></td>
<td><img src="f" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_\gamma, S^*) = 2.076$</td>
<td>$\varepsilon(\hat{S}_\gamma, S^*) = 0.6298$</td>
<td></td>
</tr>
<tr>
<td>Estimates obtained by the wavelet-based method</td>
<td><img src="g" alt="Image" /></td>
<td><img src="h" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}_w, S^*) = 0.8431$</td>
<td>$\varepsilon(\hat{S}_w, S^*) = 0.1334$</td>
<td></td>
</tr>
<tr>
<td>Estimates obtained by our method</td>
<td><img src="i" alt="Image" /></td>
<td><img src="j" alt="Image" /></td>
</tr>
<tr>
<td>$\varepsilon(\hat{S}, S^*) = 0.5624$</td>
<td>$\varepsilon(\hat{S}, S^*) = 0.07403$</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.30:** Snapshots of level set estimation simulation results corresponding to HyMapImage2 image when $K = N/2$.  

118
Conclusion

This dissertation presented methods and associated theoretical analyses for different spectral image processing tasks such as spectral image reconstruction and target detection that are useful in a wide variety of applications.

5.1 Contributions

The following are the key contributions made by this dissertation to the existing body of literature on spectral image processing:

I. Multiscale, partition-based methods for spectral image reconstruction

   (i) Development and analysis of a spatially and spectrally adaptive multi-scale intensity estimation method to estimate spectral intensities from extremely photon-limited Poisson observations.

   (ii) Extension of the intensity estimation method to Poisson inverse problems.

   (iii) Computationally efficient algorithms for Poisson spectral intensity estimation that takes $O(N^2M)$ operations to denoise spectral data of size $N \times N \times M$. 
II. Methods for spectral target detection from projection measurements

(i) Methods and associated analysis for performing whitening transformation on projection measurements that lead to a bound on the maximum tolerable background contamination.

(ii) Design of a measurement matrix that adapts to the background contamination in the underlying spectral image.

(iii) Development and analysis of a practical, nearest-neighbor based detection algorithm to detect known spectral targets from projection measurements without reconstructing the underlying spectral image.

(iv) Extension of our detection algorithm to a manifold-based target detection framework.

(v) Development and analysis of practical methods to detect spectral anomalies from projection measurements and control the FDR below some desired level without reconstructing the underlying spectral image.

III. Methods and analysis for fast level set estimation from projection measurements

(i) Development of a partition-based method to estimate level sets of a function in \( \mathbb{R}^N \) from \( K \leq N \) projection measurements without reconstructing the underlying function.

(ii) Practical and effective algorithms for set estimation that need at most \( \mathcal{O}(N(K + \log N)) \) operations.

IV. Derivation of fundamental performance bounds

(i) Proof that our multiscale, recursive dyadic partition-based spectral intensity estimation method is near-minimax optimal over the anisotropic Hölder-Besov spectral image function class.
(ii) Upper bounds on the pFDR corresponding to our nearest-neighbor based spectral target detection algorithm that offer insights on the performance of the detector as a function of (a) number of measurements, (b) signal-to-noise ratio at every spatial location, (c) similarity among potential targets of interest and (d) the prior probabilities of the targets of interest.

(iii) Tight upper bounds on the p-value of our anomaly detector decision statistic that enable effective FDR control in the anomaly detection algorithm.

5.2 Summary and future work

The partition-based, spatiotemporal denoising method presented in Chapter 2 was shown to be adaptive to varying spatial and spectral smoothness of the underlying spectral image and near-minimax optimal over an anisotropic Hölder-Besov function class. Experiments conducted on simulated and real data sets demonstrated that our algorithm can effectively extract edges and other structures from extremely photon-limited observations. These results suggest that our algorithm can be very useful in applications where binning of observations is necessary because of software limitations in estimating spectral intensities from photon-limited observations. Furthermore, our spatiotemporal denoising algorithm was successfully incorporated in an Expectation-Maximization framework to solve Poisson inverse problems.

Spectral target and anomaly detection methods presented in Chapter 3 showed that efficient target detection is possible from few specialized projection measurements collected at every spatial location. In particular, specialized projection measurements were required for performing whitening transformation on such measurements to eliminate the background contamination in the underlying spectral image. In applications such as biomedical imaging, where the projection operation is dictated by the physics of the imaging process, background elimination from projection
measurements might be very difficult. While our current problem set up does not handle such a situation, it is an interesting avenue for future work. The existing problem framework only considers projection measurements along the spectral channels. However, there exists simpler compressed sensing based camera architectures such as the singer disperser hardware architecture [30] that performs single-snapshot spectral imaging by mixing spatial and spectral information at the detectors. Though this architecture is easy to build and requires fewer optical elements, it complicates the processing techniques needed to extract relevant information from such spectral data because of the mixing in the spatial and spectral information. Some preliminary results on the extension of our target detection methods to such problem frameworks yield promising results. However, a more complete analysis on such methods is another interesting avenue for future work.

Partition-based level set estimation method discussed in Chapter 4 demonstrated that level sets can be efficiently extracted from proxy observations constructed from few projection measurements. A related work on level set estimation from direct measurements [44] shows that partition-based methods can achieve near-minimax optimality over a wide variety of function classes. This suggests that our set estimation method based on partition-based estimates should achieve similar optimality results and an exploration of such a minimax risk analysis is our ongoing work.
Appendix A

Proofs of Theorems and Lemmas

A.1 Proofs of results form Chapter 2

A.1.1 Proof of Theorem 1

For the expected squared Hellinger loss function, it can be shown [16,17] that, if we consider all density estimates in a finite or countable family $\Gamma_{M,N}$ and if the penalty corresponding to every estimate $\tilde{f} \in \Gamma_{M,N}$ satisfies the Kraft inequality, then, by applying the Li-Barron theorem [109], it can be shown that

$$\mathbb{E} \left[ \mathcal{H}^2(f, \hat{f}) \right] \leq \min_{f \in \Gamma_{M,N}} \left\{ \frac{1}{n} K\left(p_f||p_{\tilde{f}}\right) + \frac{2}{n} \text{pen}(\tilde{f}) \right\}$$  \hspace{1cm} (A.1)

where $K\left(p_f||p_{\tilde{f}}\right) \equiv \sum_y p(y|f) \log_e \left( \frac{p(y|f)}{p(y|\tilde{f})} \right)$ is the Kullback-Leibler (KL) divergence between densities $p_f$ and $p_{\tilde{f}}$ following the proof techniques in [16, 17]. Here, we use the subscript notation $p_f(y) \equiv p(y|f)$. The Li-Barron theorem shows that the performance of our estimator is bounded by the optimal tradeoff between approximation error $\left(K\left(p_f||p_{\tilde{f}}\right)\right)$ and the estimation error $\left(\text{pen}(\tilde{f})\right)$. The penalty term can be thought of as a bound on estimation errors since it is proportional to the number of degrees of freedom in an estimate.
When the observations follow a multinomial distribution, the risk function $R(f, \hat{f})$ can be bounded by [16]

$$
R(f, \hat{f}) \leq \min_{f \in \Gamma_{M,N}} \left\{ \frac{K \left( p_f \| p_{\hat{f}} \right) + 2 \text{pen}(\hat{f})}{n} \right\} \leq \min_{f \in \Gamma_{M,N}} \left\{ \frac{N^2 M \| f - \tilde{f} \|_2^2 + 2}{n} \cdot \text{pen}(\tilde{f}) \right\}.
$$

(A.2)

The squared $\ell_2$ approximation error between the discrete densities $f$ and $\tilde{f}$ can be upper bounded by the $L_2$ approximation error between the continuous densities $f$ and $\tilde{f}$ respectively:

$$
\| f - \tilde{f} \|_2^2 \leq \frac{1}{N^2 M} \| f - \hat{f} \|_{L_2}^2.
$$

(A.3)

Expressing $\| f - \tilde{f} \|_2^2$ in terms of $\| f - \hat{f} \|_2^2$ allows us to exploit the theoretical properties of the Hölder and Besov function class to bound the approximation error between $f$ and $\tilde{f}$. The approximation error between $f$ and $\tilde{f}$ can be bounded using triangle inequality as

$$
\| f - \tilde{f} \|_{L_2}^2 \leq 2 \| f - g \|_{L_2}^2 + 2 \| g - \tilde{f} \|_{L_2}^2
$$

(A.4)

where $g$ is obtained by forming a spatial partition of cuboids with sidelengths at least $1/m$, where $m \leq N$ is a dyadic number. In particular, $g$ has $O(m^{2-\gamma})$ cuboids with sidelength $1/m$ along the Hölder boundary in $f$ and $O(m)$ larger cuboids apart from the boundary. Using techniques similar to those presented in [19, 64], it can be shown that the approximation error between $f$ and $g$ decays as

$$
\| f - g \|_{L_2}^2 \leq m^{-\alpha} + m^{-\gamma} \leq m^{-\nu}
$$

(A.5)

where $\nu = \min(\alpha, \gamma)$ [64]. Given $g$, $\tilde{f}$ is obtained by finding the best $p$-piece spectral RDP of each spectrum in $g$ and fitting polynomials to every partition cell. The second term on the RHS of (A.4) can be determined using techniques described in [17, 64].
based on free-knot polynomial approximation errors [49] to yield

\[
\|g(x_1, x_2, \cdot) - \tilde{f}(x_1, x_2, \cdot)\|_{L_2} \leq p^{-\beta} + \left(\frac{p}{M}\right)^{1/2} + n^{-1/2}.
\]  
(A.6)

The approximation error in (A.6) corresponds to a single spectrum. Integrating (A.6) over the spatial domain of observations \([0, 1]^2\) we have

\[
\|g - \tilde{f}\|_{L_2}^2 \leq p^{-2\beta} + \frac{p}{M} + \frac{1}{n}
\]  
(A.7)

Substituting (A.5) and (A.7) in (A.4),

\[
\|f - \tilde{f}\|_{L_2}^2 \leq m^{-\nu} + p^{-2\beta} + \frac{p}{M} + \frac{1}{n}
\]  
(A.8)

Now we show that the penalty corresponding to the density estimates in class \(\Gamma_{M,N}\) as defined in (2.4) satisfies the Kraft inequality. Consider an estimate \(\tilde{f} \in \Gamma_{M,N}\), which, according to our definition, is piecewise constant in the two spatial dimensions and piecewise polynomial in the spectral dimension. In a tree representation, it can be represented as a quadtree having binary splits in each of its leaves. We arrive at the penalty for the estimate \(\tilde{f}\) by computing the number of bits needed to uniquely represent \(\tilde{f}\). The structure of a quadtree with \(k\) leaves can be uniquely encoded using at most \(4k/3\) bits [17]. Each of the \(k\) leaves of the quadtree is associated with a binary tree which can also be encoded using a prefix code as described in [17]. In particular, a binary tree with \(p\) leaves consists of \(2p - 1\) nodes, and hence can be encoded uniquely by at most \(2p - 1\) bits, which can be verified by induction [17]. Let \(P\) be the spatial partition, and for each \(c \in P\), let \(p_c\) be the number of cells in the spectral RDP. Then the number of bits required to encode the structure of the spatiotemporal RDP \(Q\) is

\[
\frac{4}{3}k + \sum_{c \in P}(2p_c - 1) \leq \frac{4}{3}k + 2|Q| \leq \frac{10}{3}|Q|.
\]
Assuming that each polynomial coefficient is quantized to one of $\sqrt{n}$ levels, $\log_2 \sqrt{n}$ bits are needed to encode each coefficient in the partition $Q$, and the total number of bits required to prefix encode the estimate is $|Q| \left( \frac{10}{3} + \frac{r}{2} \log_2 n \right)$. Thus the penalty in (2.4) corresponds to a prefix codelength and is guaranteed to satisfy

$$\sum_{\tilde{f} \in \Gamma_{M,N}} 2^{-|Q(\tilde{f})| \left( \frac{10}{3} + \frac{r}{2} \log_2 n \right)} \leq 1,$$

or

$$\sum_{\tilde{f} \in \Gamma_{M,N}} e^{-|Q(\tilde{f})| \left( \frac{10}{3} + \frac{r}{2} \log_2 n \right) \log_e 2} \leq 1,$$

as desired.

The approximation error in (A.8) assumed $k = O \left( m^{2-\gamma} \right)$ cells in the spatial partition and a maximum of $p$ cells in each spectral partition, for a total of at most $kp$ cells. It further assumed that the spectral approximation was a free-knot piecewise polynomial. However, in the RDP construction, the knots are restricted to lie on dyadic interval endpoints. A $p$-piece piecewise polynomial can be subdivided into a $(p \log_2 M)$-piece piecewise polynomial with the same approximation error but with knots on dyadic interval endpoints. Thus the RDP corresponding to the approximation error in (A.8) has at most $kp \log_2 M$ cells, so its penalty is bounded by

$$\text{pen}(\tilde{f}) \leq kp \log_2 M \left( \frac{10}{3} + \frac{r}{2} \log_2 n \right) \log_e 2,$$

$$< kp \log_2 M \left( \frac{10}{3} + \frac{r}{2} \right) \log_2 n \log_e 2 = kp \log_2 M \left( \frac{10}{3} + \frac{r}{2} \right) \log_e n. \quad (A.9)$$

Applying (A.3), (A.8) and (A.9) to (A.2), the risk function can be rewritten as follows:

$$R(f, \tilde{f}) \leq \min_{m,p} \left\{ \left( m^{-\nu} + p^{-2\beta} + \frac{p}{M} + \frac{1}{n} \right) + \left( \frac{10}{3} + \frac{r}{2} \right) m^{2-\gamma} p \log_2 M \log_e n \right\}$$

126
The values of $m$ and $p$ that minimize the risk function given above can be shown to be

$$m = C'(\frac{n}{\log_2 M \log_e n})^{\frac{2\beta}{2\beta+2-\gamma+\nu}}$$
and
$$p = C'(\frac{n}{\log_2 M \log_e n})^{\frac{\nu}{2\beta+2-\gamma+\nu}}$$

where $C'$ is a constant independent of $n$, $N$, and $M$ and dependent on the smoothness parameters $\alpha$, $\beta$ and $\gamma$. Thus, if we collect the measurements on a grid of size $N \times N \times M$, where $N \geq m$ and $M \geq p$ and estimate the true intensity $f$ by the estimation procedure given in Sec. 2.2, then the risk function between the true and the estimated density is upper bounded by,

$$R(f, \hat{f}) \leq (\frac{n}{\log_2 M \log_e n})^{\frac{-2\beta\nu}{2\beta+2-\gamma+\nu}}$$

which is within a log factor of the lower bound when $\gamma = 1$.

A.2 Proofs of results from Chapter 3

A.2.1 Proof of Theorem 3

Using linear algebra and matrix theory, it is possible to show that if $B = I - A\Sigma_b A^T$ is positive definite, then

$$\Phi = \sigma B^{-1/2}A$$  \hspace{1cm} (A.10)

satisfies (3.3).$^1$ In particular, we can substitute (A.10) in (3.3) to verify that the proposed construction of $\Phi$ satisfies (3.3). Observe that $C_\Phi = (\Phi\Sigma_0\Phi^T + \sigma^2 I)^{-1/2}$

$^1$ We would like to thank Prof. Roummel Marcia for fruitful discussions related to this point.
can be written in terms of (A.10) as follows:

\[
C_{\Phi} = \left( \left[ \sigma B^{-1/2} A \right] \Sigma_b \left[ \sigma B^{-1/2} A \right]^T + \sigma^2 I \right)^{-1/2}
\]

\[
= \left( \sigma^2 B^{-1/2} \left( A \Sigma_b A^T \right) \left( B^{-1/2} \right)^T + \sigma^2 I \right)^{-1/2}
\]

\[
= \left( \sigma^2 B^{-1/2} \left( I - B \right) \left( B^{-1/2} \right)^T + \sigma^2 I \right)^{-1/2}
\]

(A.11)

\[
= \left( \sigma^2 B^{-1} \right)^{-1/2} = \sigma^{-1} B^{1/2}
\]

(A.12)

where (A.11) follows from the definition of \( B \) and (A.12) follows from the fact that \( B \) is symmetric and positive definite. If \( B \) is positive definite, then \( B^{-1} \) is positive definite as well and can be decomposed as \( B^{-1} = \left( B^{-1/2} \right)^T \left( B^{-1/2} \right) \), where the matrix square root \( B^{-1/2} \) is symmetric and positive definite. By substituting (A.12) and (A.10) in (3.3), we have

\[
C_{\Phi} \Phi = \sigma^{-1} B^{1/2} \sigma B^{-1/2} A = A.
\]

A sufficient condition for \( B \) to be positive definite can be derived as follows.

To ensure positive definiteness of \( B \), we must have

\[
x^T B x = x^T x - x^T \left( A \Sigma_b A^T \right) x > 0 \quad (A.13)
\]

for any nonzero \( x \in \mathbb{R}^K \). Note that since \( \Sigma_b \) is positive semidefinite, \( x^T \left( A \Sigma_b A^T \right) x \geq 0 \). However, the right hand side of (A.13) is > 0 only if the spectral norm of \( A \Sigma_b A^T \) is < 1, since \( x^T \left( A \Sigma_b A^T \right) x \leq \|x\|^2 \cdot \|A \Sigma_b A^T\| \). The norm of \( A \Sigma_b A^T \) is in turn bounded above by

\[
\|A \Sigma_b A^T\| \leq \|A\| \|\Sigma_b\| \|A^T\| = \|A\|^2 \|\Sigma_b\| = \|A\|^2 \lambda_{\text{max}}
\]

since \( \|A\| = \|A^T\| \) and \( \|\Sigma_b\| = \lambda_{\text{max}} \), where \( \lambda_{\text{max}} \) is the largest eigenvalue of \( \Sigma_b \). To ensure \( \|A \Sigma_b A^T\| < 1 \), \( \|A\|^2 \lambda_{\text{max}} \) has to be < 1, which leads to the result of Theorem 3.

A.2.2 Proof of Theorem 4

The proof of Theorem 4 adapts the proof techniques from [36] to nonidentical independent hypothesis tests. We begin by expanding the pFDR definition in (3.10) as
follows:
\[
pFDR^{(j)}(\Gamma) = \sum_{k=1}^{M} \mathbb{E} \left[ \frac{V(\Gamma)}{R(\Gamma)} \left| R(\Gamma) = k \right. \right] \times \mathbb{P} \left( R(\Gamma) = k \left| R(\Gamma) > 0 \right. \right).
\]

Observe that \( R(\Gamma) = k \) implies that there exists some subset \( S_k = \{u_1, \ldots, u_k\} \subseteq \{1, \ldots, M\} \) of size \( k \) such that \( y_{u_\ell} \in \Gamma_{u_\ell}^{(j)} \) for \( \ell = 1, \ldots, k \) and \( y_i \not\in \Gamma_i^{(j)} \) for all \( i \not\in S_k \). To simplify the notation, let \( \Lambda_{S_k} = \prod_{u \in S_k} \Gamma_u^{j} \times \prod_{\ell \not\in S_k} \tilde{\Gamma}_\ell^{(j)} \), where \( \tilde{\Gamma}_\ell^{(j)} \) is the complement of \( \Gamma_\ell^{(j)} \), denote the significance region that corresponds to set \( S_k \), and \( T = (y_1, \ldots, y_M) \) be a set of test statistics corresponding to each hypothesis test. Considering all such subsets we have
\[
pFDR^{(j)}(\Gamma) = \sum_{k=1}^{M} \sum_{S_k} \mathbb{E} \left[ \frac{V(\Gamma)}{k} \left| T \in \Lambda_{S_k} \right. \right] \times \mathbb{P} \left( T \in \Lambda_{S_k} \left| R(\Gamma) > 0 \right. \right).
\] (A.14)

By plugging in the definition of \( V(\{\Gamma_i\}) \) from (3.11), we have
\[
\mathbb{E} \left[ V(\Gamma) \left| T \in \Lambda_{S_k} \right. \right] = \mathbb{E} \left[ \sum_{i=1}^{M} \mathbb{I} \{y_i \in \Gamma_i^{(j)}\} \mathbb{I} \{H_i^{(j)} = 0\} \left| T \in \Lambda_{S_k} \right. \right]
\]
\[
= \sum_{\ell=1}^{k} \mathbb{E} \left[ \mathbb{I} \{H_{u_\ell}^{(j)} = 0\} \left| y_{u_\ell} \right. \right] = \sum_{\ell=1}^{k} \mathbb{P} \left( H_{u_\ell}^{(j)} = 0 \left| y_{u_\ell} \in \Gamma_{u_\ell}^{(j)} \right. \right) \quad (A.15)
\]
for all \( u_\ell \in S_k \) since the tests are independent of each other. The posterior probability
\[
\mathbb{P} \left( H_i^{(j)} = 0 \left| y_i \in \Gamma_i^{(j)} \right. \right) \quad \text{for the } i^{th} \text{ hypothesis test can be expanded using Bayes’ rule as}
\]
\[
\mathbb{P} \left( H_{0i}^{(j)} \left| y_i \in \Gamma_i^{(j)} \right. \right) = \frac{\mathbb{P} \left( y_i \in \Gamma_i^{(j)} \left| H_{0i} \right. \right) \mathbb{P} \left( H_{0i}^{(j)} \right)}{\mathbb{P} \left( y_i^{(j)} \in \Gamma_i^{(j)} \right)}
\]
\[
= \frac{\mathbb{P} \left( \hat{f}_i \neq f_i^{(j)} \left| f_i^{*} = f_i^{(j)} \right. \right) \mathbb{P} \left( f_i^{*} = f_i^{(j)} \right)}{\mathbb{P} \left( \hat{f}_i \neq f_i^{(j)} \right)},
\] (A.16)
where \( \hat{f}_i = \arg \max_{f \in \mathcal{D}} \mathbb{P} (f_i^* = f \mid y_i, \alpha_i, A) \). To upper bound the numerator of (A.16), consider the probability of misclassification given by \((P_e)_i = \mathbb{P} (\hat{f}_i \neq f_i^*)\) where \( f_i^* = f^{(j)} \in \mathcal{D} \), which can be expanded as follows:

\[
(P_e)_i = \mathbb{P} (\hat{f}_i \neq f_i^*) = \sum_{\ell=1}^{m} \mathbb{P} (\hat{f}_i \neq f_i^* \mid f_i^* = f^{(\ell)}) \mathbb{P} (f_i^* = f^{(\ell)})
\]

\[
\equiv \sum_{\ell=1}^{m} \mathbb{P} (\hat{f}_i \neq f^{(\ell)}) \mathbb{P} (f_i^* = f^{(\ell)})
\]

\[
\geq \mathbb{P} (\hat{f}_i \neq f^{(j)}) \mathbb{P} (f_i^* = f^{(j)})
\]

(A.17)

The denominator term in (A.16) can be expanded as follows:

\[
\mathbb{P} (\hat{f}_i \neq f^{(j)}) = \mathbb{P} (\hat{f}_i \neq f^{(j)} \mid f_i^* = f^{(j)}) \mathbb{P} (f_i^* = f^{(j)})
\]

\[
+ \mathbb{P} (\hat{f}_i \neq f^{(j)} \mid f_i^* \neq f^{(j)}) \mathbb{P} (f_i^* \neq f^{(j)})
\]

Observe that \( \mathbb{P} (\hat{f}_i \neq f^{(j)} \mid f_i^* = f^{(j)}) \) is nonnegative, and

\[
\mathbb{P} (\hat{f}_i \neq f^{(j)} \mid f_i^* \neq f^{(j)}) = \mathbb{P} (\hat{f}_i \in \mathcal{D} \setminus f^{(j)} \mid f_i^* \neq f^{(j)})
\]

\[
\geq \mathbb{P} (\hat{f}_i = f_i^* \mid f_i^* \neq f^{(j)}) = 1 - \mathbb{P} (\hat{f}_i \neq f_i^* \mid f_i^* \neq f^{(j)})
\]

\[
= 1 - \frac{\mathbb{P} (\hat{f}_i \neq f_i^*, f_i^* \neq f^{(j)})}{\mathbb{P} (f_i^* \neq f^{(j)})} \geq 1 - \frac{(P_e)_i}{1 - p^{(j)}}
\]

(A.18)

Thus

\[
\mathbb{P} (\hat{f}_i \neq f^{(j)}) \geq \left( 1 - \frac{(P_e)_i}{1 - p^{(j)}} \right) (1 - p^{(j)}) = 1 - p^{(j)} - (P_e)_i.
\]

Substituting (A.17) and (A.18) in (A.16),

\[
\mathbb{P} (H_{0i}^{(j)} \mid y_i \in \Gamma_i^{(j)}) \leq \frac{(P_e)_i}{1 - p^{(j)} - (P_e)_i} \leq \frac{\max_e (P_e)}{1 - p^{(j)} - \max_e (P_e)}.
\]

(A.19)
By substituting (A.19) in (A.14) and (A.15) we have:

\[ p_{\text{FDR}}(j)(\Gamma) \leq \sum_{k=1}^{M} \sum_{S_k} \frac{1}{k} \left( \sum_{\ell=1}^{k} \frac{(P_e)_{\max}}{1 - p^{(j)} - (P_e)_{\max}} \right) \times \mathbb{P}(T \in \Lambda_{S_k} \mid R(\Gamma) > 0) \]

\[ = \frac{(P_e)_{\max}}{1 - p^{(j)} - (P_e)_{\max}} \sum_{k=1}^{M} \sum_{S_k} \mathbb{P}(T \in \Lambda_{S_k} \mid R(\Gamma) > 0) \]

\[ \leq \frac{(P_e)_{\max}}{1 - p^{(j)} - (P_e)_{\max}} \]

since \( \sum_{k=1}^{M} \sum_{S_k} \mathbb{P}(T \in \Lambda_{S_k} \mid R(\Gamma) > 0) \leq 1 \). The result of Theorem 4 is obtained by finding an upper bound on the worst-case pFDR given by

\[ p_{\text{FDR}}_{\max} = \max_{j \in \{1, \ldots, m\}} p_{\text{FDR}}(j)(\Gamma) \]

\[ \leq \max_{j \in \{1, \ldots, m\}} \frac{(P_e)_{\max}}{1 - p^{(j)} - (P_e)_{\max}} = \frac{(P_e)_{\max}}{1 - p_{\max} - (P_e)_{\max}} \]

where \( p_{\max} = \max_{\ell \in \{1, \ldots, m\}} p^{(\ell)} \).

A.2.3 Proof of Corollary 5

By construction, the entries of \( A \) are drawn from \( \mathcal{N}(0, 1/K) \). It has been shown [110] that if \( G \) is a \( K \times N \) random matrix with i.i.d. zero mean subgaussian entries with unit variance, then, with probability at least \( 1 - \exp(-c(K+N)) \), the following holds:

\[ \|G\| \leq c\left(\sqrt{K} + \sqrt{N}\right) \]

where \( c \) is an absolute constant. Applying this result to \( A \) yields \( \|A\| \leq c\left(\sqrt{N/K} + 1\right) \) with high probability. This result, together with the result of Theorem 3 yields Corollary 5.

A.2.4 Proof of Corollary 6

We prove Corollary 6 by bounding the worst-case misclassification probability \((P_e)_{\max}\) corresponding to a Gaussian sensing matrix \( A \) and applying the result of Theorem 4.
When the entries of $A$ are drawn from $\mathcal{N}(0, 1/K)$, the lemma below gives an upper bound on the misclassification probability $(P_e)_i = \mathbb{P} \left( \hat{f}_i \neq f^*_i \right)$ at the $i$th spatial location, which leads to an upper bound on the worst-case misclassification probability:

**Lemma 9.** *(Compressive classification error [101])* Consider the problem of classifying a signal of interest $f^*_i \in \mathcal{D} = \{f^{(1)}, \ldots, f^{(m)}\}$ at the $i$th spatial location to one of $m$ known target classes by making observations of the form in (3.2), given the knowledge of the spectral dictionary $\mathcal{D}$, and the sensing matrix $A$. If the estimate $\hat{f}_i$ at the $i$th spatial location is obtained according to (3.14), then

$$
(P_e)_i = \mathbb{P} \left( \hat{f}_i \neq f^*_i \right) \leq \frac{1 - p_{\min}}{p_{\min}} \left( 1 + \frac{\alpha_i^2 d_{\min}^2}{4K} \right)^{-\frac{K}{2}} 
$$

(A.20)

where $d_{\min} = \min_{f^{(i)}, f^{(j)} \in \mathcal{D}, i \neq j} \|f^{(i)} - f^{(j)}\|$.

This lemma is very similar to Theorem 1 in [101], which was proved using a combination of a maximum a posteriori (MAP) framework and the moment generating function for Gaussian random variables. Our innovation consists of accounting for a priori non-equiequiprobable dictionary elements; this is a straightforward modification of the proof in [101] and is not presented here. The worst-case misclassification probability can be obtained directly from (A.20) as follows:

$$
(P_e)_{\text{max}} = \max_{i \in \{1, \ldots, M\}} \left[ \frac{1 - p_{\min}}{p_{\min}} \left( 1 + \frac{\alpha_i^2 d_{\min}^2}{4K} \right)^{-\frac{K}{2}} \right]
$$

$$
= \frac{1 - p_{\min}}{p_{\min}} \left( 1 + \frac{\alpha_{\min}^2 d_{\min}^2}{4K} \right)^{-\frac{K}{2}}
$$

(A.21)

where $\alpha_{\min} = \min_{i \in \{1, \ldots, M\}} \alpha_i$. The result in (3.16) is obtained by substituting (A.21) in (3.13), and simplifying the resulting expression.
A.2.5 Proof of Theorem 7

We first prove this theorem assuming that \(\{\alpha_i\}\) are known and later extend to the case where \(\{\hat{\alpha}_i\}\) are estimated form the observations. Let \(\tilde{f}_i = \arg\min_{f \in \mathcal{D}} \|f_i^* - f\|\).

The p-value expression in (3.19) can be expanded as follows:

\[
p_i = \mathbb{P}\left(\bar{d}_i \geq d_i \mid \mathcal{H}_{0i}\right)
= \mathbb{P}\left(\min_{f \in \mathcal{D}} \|\alpha_i A(f_i^* - f) + n\| \geq d_i \mid \mathcal{H}_{0i}\right)
\leq \mathbb{P}\left(\|\alpha_i A(f_i^* - \tilde{f}_i) + n\| \geq d_i \mid \mathcal{H}_{0i}\right)
= \mathbb{P}\left(\|\alpha_i A(f_i^* - \tilde{f}_i) + n\|^2 \geq d_i^2 \mid \mathcal{H}_{0i}\right). \tag{A.22}
\]

Note that \(\|\alpha_i A(f_i^* - \tilde{f}_i) + n\|^2\) is a noncentral \(\chi^2\) random variable with \(K\) degrees of freedom and a noncentrality parameter \(\nu_i = \|\alpha_i A(f_i^* - \tilde{f}_i)\|^2\). Thus (A.22) can be written in terms of a noncentral \(\chi^2\) CDF \(F(d_i^2; K, \nu_i)\) with parameter \(d_i^2\). The upper and lower bounds on \(\nu_i\) can be obtained using the properties of the projection matrix \(A\). Applying (3.20), we see that

\[
\alpha_i^2(1 - \epsilon)^2\|f_i^* - \tilde{f}_i\|^2 \leq \nu_i \leq \alpha_i^2(1 + \epsilon)^2\|f_i^* - \tilde{f}_i\|^2
\]

with high probability. Thus,

\[
p_i \leq 1 - \mathbb{P}\left(\|\alpha_i A(f_i^* - \tilde{f}_i) + n\|^2 \leq d_i^2 \mid \mathcal{H}_{0i}\right) \tag{A.23}
= 1 - F(d_i^2; K, \nu_i)
\leq 1 - F(d_i^2; K, \alpha_i^2(1 + \epsilon)^2\|f_i^* - \tilde{f}_i\|^2)
\leq 1 - F(d_i^2; K, \alpha_i^2(1 + \epsilon)^2\tau^2)
\]

since \(\|f_i^* - f\| \leq \tau\) for all \(f \in \mathcal{D}\) under \(\mathcal{H}_{0i}\).

When \(\{\alpha_i\}\) are estimated from the observations such that \(\{\hat{\alpha}_i\}\) satisfy (3.21), we
can write the p-value expression in (A.23) as follows:

\[
p_i \leq 1 - F \left( d_i^2; K, \| A \left( \alpha_i f_i^* - \tilde{\alpha}_i \tilde{f}_i \right) \| \right)^2
\]

\[
\leq 1 - F \left( d_i^2; K, (1 + \epsilon)^2 \tilde{\alpha}_i^2 \| \frac{\alpha_i}{\tilde{\alpha}_i} f_i^* - \tilde{f}_i \| \right)^2
\]

(A.24)

where (A.24) is due to the distance preservation property of \( A \) given in (3.20).

Observe that \( \| \frac{\alpha_i}{\tilde{\alpha}_i} f_i^* - \tilde{f}_i \|^2 \) can be upper bounded as shown below:

\[
\left\| \frac{\alpha_i}{\tilde{\alpha}_i} f_i^* - \tilde{f}_i \right\|^2 = \left\| \left( \frac{\alpha_i}{\tilde{\alpha}_i} - 1 \right) f_i^* + f_i^* - \tilde{f}_i \right\|^2
\]

\[
\leq \left( \left\| \left( \frac{\alpha_i}{\tilde{\alpha}_i} - 1 \right) f_i^* \right\| + \left\| f_i^* - \tilde{f}_i \right\| \right)^2
\]

(A.25)

\[
= \left( \left\| \frac{\alpha_i}{\tilde{\alpha}_i} - 1 \right\| + \left\| f_i^* - \tilde{f}_i \right\| \right)^2
\]

(A.26)

where (A.25) is due to the triangle inequality, (A.26) comes from the assumption that \( \| f_i^* \| = 1 \), and the last inequality is due to (3.21). By applying this result to (A.24) and exploiting the fact that \( \| f_i^* - f \| \leq \tau \) under \( \mathcal{H}_0 \) for some \( f \in \mathcal{D} \), we have

\[
p_i \leq 1 - F \left( d_i^2; K, (1 + \epsilon)^2 \tilde{\alpha}_i^2 \left( \zeta + \left\| f_i^* - \tilde{f}_i \right\| \right)^2 \right)
\]

\[
\leq 1 - F \left( d_i^2; K, (1 + \epsilon)^2 \tilde{\alpha}_i^2 \left( \zeta + \tau \right)^2 \right).
\]

A.3 Proofs of results from Chapter 4

A.3.1 Proof of Theorem 8

Let us begin by bounding \( T_1 \) and \( T_2 \) in (4.6) separately. Let \( \tilde{p}_L = \sum_{i \in L} \frac{1}{N} \). From the statistics of \( z \), we can bound \( T_1 = \frac{1}{N} \sum_{i=1}^{N} \left( \mathbb{E} \left[ z_i \right] - f_i \right) \left[ \mathbb{I}_{\{\ell(L) = 1\}} - \mathbb{I}_{\{\ell(L) = 0\}} \right] \mathbb{I}_{\{i \in L\}} \) as
follows:

$$T_1 \leq \frac{1}{N} \sum_{i,j:j \neq i} |f_j|\|\langle A^{(i)}, A^{(j)} \rangle\| \left[ \|I\{\ell(L)=1\} - I\{\ell(L)=0\}\| \right] I\{i \in L\}$$

$$\leq \frac{\mu(A)}{N} \sum_{i \in L} \sum_{j=1,j \neq i}^N |f_j| = \frac{\mu(A)}{N} \sum_{i \in L} \left( \sum_{j=1}^N |f_j| - |f_i| \right)$$

$$\leq \mu(A)\|f\|_1 - \frac{\mu(A)}{N} \sum_{i \in L} |f_i|$$

(A.27)

where the third inequality is due to the fact that $$\|I\{\ell(L)=1\} - I\{\ell(L)=0\}\| = 1$$ and $$\|\langle A^{(i)}, A^{(j)} \rangle\| \leq \mu(A)$$ for all $$j \neq i$$.

We can bound $$T_2$$, which is a function of $$z$$, using McDiarmid’s inequality by observing that $$z_i = \sum_{j=1}^K a_{j,i}y_j$$ and thus $$T_2$$ can be written as a function of the independent random variables $$\{y_k\}$$. Specifically,

$$T_2 = \frac{1}{N} \sum_{i \in L} (z_i - \mathbb{E}[z_i]) \left[ I\{\ell(L)=1\} - I\{\ell(L)=0\} \right]$$

$$= \sum_{j=1}^K y_jw_j(L) - \mathbb{E} \left[ \sum_{j=1}^K y_jw_j(L) \right]$$

where

$$w_j(L) = \sum_{i \in L} \left[ I\{\ell(L)=1\} - I\{\ell(L)=0\} \right] \frac{a_{j,i}}{N}.$$  \hspace{1cm} (A.28)

We can formally state the McDiarmid’s inequality as follows: Suppose that $$y_1, \ldots, y_K$$ are independent random variables and that a function $$g(y_1, \ldots, y_K)$$ satisfies

$$\sup_{\bar{y}_1, \ldots, \bar{y}_K} |g(\bar{y}_1, \ldots, \bar{y}_K) - g(\bar{y}^{(p-1)}, \bar{y}_p, \bar{y}_{p+1}, \ldots, \bar{y}_K)| \leq c_p$$ \hspace{1cm} (A.29)

where $$p = 1, \ldots, K$$, $${\bar{y}_i}_{i=1}^K$$ are the values that the independent random variables $${y_i}_{i=1}^K$$ take and $$\bar{y}^{(p-1)} = {\bar{y}_1, \ldots, \bar{y}_{p-1}}$$. Then, for any $$\epsilon > 0$$

$$\mathbb{P}(g(y_1, \ldots, y_K) - \mathbb{E}[g(y_1, \ldots, y_K)] \geq \epsilon) \leq \exp \left( \frac{-2\epsilon^2}{\sum_{p=1}^K c_p^2} \right).$$ \hspace{1cm} (A.30)
We can thus bound $T_2$ using the McDiarmid’s inequality as long as $g(y_1, \ldots, y_K) = \sum_{j=1}^{K} y_j w_j(L)$ satisfies (A.29). Since the entries of noise $n$ are bounded, it can be shown that the function $g(y_1, \ldots, y_K)$ satisfies the bounded differences property \[107\] as follows:

\[
\sup_{\tilde{y}_1, \ldots, \tilde{y}_K, \tilde{y}_p} \left| g(\tilde{y}_1, \ldots, \tilde{y}_K) - g(\tilde{y}_1^{(p-1)}, \tilde{y}_p', \tilde{y}_{p+1}, \ldots, \tilde{y}_K) \right|
\]

\[= \left| w_p(L) \right| \sup_{\tilde{y}_1, \ldots, \tilde{y}_K, \tilde{y}_p} |\tilde{y}_p - \tilde{y}_p'| \equiv \left| w_p(L) \right| \sup_{\tilde{n}_1, \ldots, \tilde{n}_K, \tilde{n}_p} \left| \tilde{n}_p - \tilde{n}_p' \right| = \left| w_p(L) \right| |c_u - c_\ell|. \tag{A.31}\]

where $\{\tilde{n}_i\}_{i=1}^{K}$ are the values that the random variables $\{n_i\}_{i=1}^{K}$ take. Now, using McDiarmid’s inequality we arrive at the following result:

\[
\mathbb{P}(T_2 \geq \epsilon) \leq \exp \left( \frac{-2\epsilon^2}{\sum_{p=1}^{K} |w_p(L)|^2 |c_u - c_\ell|^2} \right). \tag{A.32}\]

Note that we can expand $\sum_{p=1}^{K} |w_p(L)|^2$ in (A.32) from (A.28) as

\[
\sum_{p=1}^{K} |w_p(L)|^2 = \sum_{p=1}^{K} \left[ \mathbb{I}(\ell(L)=1) - \mathbb{I}(\ell(L)=0) \right] \sum_{i \in L} \frac{a_{p,i}}{N} \right|^2
\]

\[= \sum_{p=1}^{K} \left[ \mathbb{I}(\ell(L)=1) - \mathbb{I}(\ell(L)=0) \right]^2 \sum_{i \in L} \frac{a_{p,i}}{N} \right|^2
\]

\[= \sum_{p=1}^{K} \left\| \sum_{i \in L} \frac{a_{p,i}}{N} \right\|^2 = \sum_{p=1}^{K} \left\| \sum_{i \in L} \frac{a_{p,i}}{N} \right\| \sum_{j \in L} \frac{a_{p,j}}{N} \right| \]

\[= \sum_{i \in L} \sum_{j \in L} \frac{1}{N^2} \left( \sum_{p=1}^{K} a_{p,i} a_{p,j} \right) = \frac{1}{N^2} \sum_{i \in L} \sum_{j \in L} \langle A^{(i)} A^{(j)} \rangle. \tag{A.33}\]
By substituting (A.33) in (A.32) and by equating the right hand side of (A.32) to \( \delta_L \in (0, 1/2) \) and solving for \( \epsilon \), we can show that, with probability at least \( 1 - \delta_L \),

\[
T_2 \leq \sqrt{\frac{\log(1/\delta_L) |c_u - c_\ell|^2 \sum_{i,j \in L} \langle A^{(i)} A^{(j)} \rangle}{2N^2}}. \tag{A.34}
\]

Applying the bounds in (A.27) and (A.34) to (4.6) we can see that with probability at least \( 1 - \delta_L \), the following holds:

\[
R(L) - \hat{R}(L) \leq \left( \mu(A) \hat{p}_L \|f\|_1 - \frac{\mu(A)}{N} \sum_{i \in L} |f_i| \right) + \sqrt{\frac{\log(1/\delta_L) |c_u - c_\ell|^2 \sum_{i,j \in L} \langle A^{(i)} A^{(j)} \rangle}{2N^2}}.
\]

Thus for a given \( S \in S_M \), the risk difference \( R(S) - \hat{R}(S) \) is upper bounded by summing the bound corresponding to each leaf separately. Since \( \sum_{L \in \pi(S)} \hat{p}_L = 1 \) and \( \sum_{L \in \pi(S)} \sum_{i \in L} |f_i| = \|f\|_1 \) we have

\[
R(S) - \hat{R}(S) \leq \mu(A) \left( \frac{N - 1}{N} \right) \|f\|_1 + \sum_{L \in \pi(S)} \sqrt{\frac{\log(1/\delta_L) |c_u - c_\ell|^2 \sum_{i,j \in L} \langle A^{(i)} A^{(j)} \rangle}{2N^2}}.
\]

with high probability. If we let \( \delta_L = \delta 2^{-\left\lfloor \|L\| + 1 \right\rfloor} \) where \( \|L\| \) is the number of bits required to uniquely encode the position of leaf \( L \), then it is straightforward to follow the proof of Lemma 2 in [44] to show that the bound above holds for every \( S \in S \), which leads to the result of Theorem 8.
Bibliography


Biography

Kalyani Krishnamurthy was born in Chennai, India, on Mar 30, 1981 to Banumathy and Krishnamurthy. She was raised in the beautiful temple town of Kumbakonam, India, and earned her bachelors degree in Electronics and Communication Engineering from Shanmugha College of Engineering (now SASTRA University), Thanjavur, India. After her undergraduate education, she worked as a Graduate Engineering Trainee at the Embedded Systems division of Larsen & Toubro, Mysore, India, from Aug 2002 to Sep 2004.

She moved to the United States post her marriage with Shivakumar Vasanth, who was then a graduate student at the Kent State University, OH. From Aug 2005 to Nov 2005, she was a visiting researcher at the Vision research lab, UCSB, working with Prof. Manjunath and Dr. Baris Sumengen on project Cortina, a large-scale content-based image retrieval system.

She started her PhD at Duke University in Jan 2006, under the guidance of Prof. Rebecca Willett. While at Duke, she earned the Charles Rowe Memorial award for being the outstanding graduate teaching assistant in Fall 2008. She is currently a Preparing Future Faculty (PFF) fellow at Duke for the academic year 2010-2011. Her research interests include statistical signal processing, spectral imaging and compressed sensing.