Sparse Inverse Problems: The Mathematics of Precision Measurement

by

Geoffrey Robert Schiebinger

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Benjamin Recht, Chair
Professor Martin Wainwright
Professor Adityanand Guntuboyina
Professor Laura Waller

Spring 2016

Sparse Inverse Problems: The Mathematics of Precision Measurement

Abstract

Sparse Inverse Problems: The Mathematics of Precision Measurement

by

Geoffrey Robert Schiebinger

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Benjamin Recht, Chair

The interplay between theory and experiment is the key to progress in the natural sciences. This thesis develops the mathematics of distilling knowledge from measurement. Specifically, we consider the inverse problem of recovering the input to a measurement apparatus from the observed output. We present separate analyses for two different models of input signals. The first setup is superresolution. Here, the input is a collection of continuously parameterized sources, and we observe a weighted superposition of signals from all of the sources. The second setup is unsupervised classification. The input is a collection of categories, and the output is an unlabeled set of objects from the different categories. In Chapter 1 we introduce these measurement modalities in greater detail and place them in a common framework.

Chapter 2 provides a theoretical analysis of diffraction-limited superresolution, demonstrating that arbitrarily close point sources can be resolved in ideal situations. Precisely, we assume that the incoming signal is a linear combination of M shifted copies of a known waveform with unknown shifts and amplitudes, and one only observes a finite collection of evaluations of this signal. We characterize properties of the base waveform such that the exact translations and amplitudes can be recovered from $2M+1$ observations. This recovery can be achieved by solving a weighted version of basis pursuit over a continuous dictionary. Our analysis shows that ℓ_1-based methods enjoy the same separation-free recovery guarantees as polynomial root finding techniques such as Prony's method or Vetterli's method for signals of finite rate of innovation. Our proof techniques combine classical polynomial interpolation techniques with contemporary tools from compressed sensing.

In Chapter 3 we propose a variant of the classical conditional gradient method (CGM) for superresolution problems with differentiable measurement models. Our algorithm combines nonconvex and convex optimization techniques: we propose global conditional gradient steps alternating with nonconvex local search exploiting the differentiable observation model. This hybridization gives the theoretical global optimality guarantees and stopping conditions of convex optimization along with the performance and modeling flexibility associated with nonconvex optimization. Our experiments demonstrate that our technique achieves state-of-the-art results in several applications.

Chapter 4 focuses on unsupervised classification. Clustering of data sets is a standard problem in many areas of science and engineering. The method of spectral clustering is based on embedding the data set using a kernel function, and using the top eigenvectors of the normalized Laplacian to recover the connected components. We study the performance of spectral clustering in recovering the latent labels of i.i.d. samples from a finite mixture of nonparametric distributions. The difficulty of this label recovery problem depends on the overlap between mixture components and how easily a mixture component is divided into two nonoverlapping components. When the overlap is small compared to the indivisibility of the mixture components, the principal eigenspace of the population-level normalized Laplacian operator is approximately spanned by the square-root kernelized component densities. In the finite sample setting, and under the same assumption, embedded samples from different components are approximately orthogonal with high probability when the sample size is large. As a corollary we control the fraction of samples mislabeled by spectral clustering under finite mixtures with nonparametric components.

To my loving family:

Dad, *for igniting my scientific curiosity;*
Mom, *for showing me the path to success;*
Jonny, *my beloved brother, for keeping me on top of my game;*
Elina, *my loving wife, for our lifetime of love, learning, and adventure ahead.*

Contents

Contents iv

1 Introduction 1
- 1.1 Superresolution 3
- 1.2 Spectral clustering 6

2 Superresolution without Separation 9
- 2.1 Introduction 9
- 2.2 Proofs 17
- 2.3 Numerical experiments 29
- 2.4 Conclusions and Future Work 34

3 The Alternating Descent Conditional Gradient Method 36
- 3.1 Introduction 36
- 3.2 Example applications 41
- 3.3 Conditional gradient method 44
- 3.4 Related work 49
- 3.5 Theoretical guarantees 50
- 3.6 Numerical results 53
- 3.7 Conclusions and future work 58

4 The Geometry of Kernelized Spectral Clustering 60
- 4.1 Introduction 60
- 4.2 Background and problem set-up 62
- 4.3 Analysis of the normalized Laplacian embedding 64
- 4.4 Proofs 75
- 4.5 Discussion 83
- 4.6 Supplementary proofs for Theorem 1 84
- 4.7 Supplementary proofs for Theorem 2 89
- 4.8 Background 94
- 4.9 List of symbols 97

Bibliography 98

Acknowledgments

First and foremost, I would like to thank my advisor, Ben Recht, for introducing me to a rich and interesting area of applied mathematics. Our work together on superresolution has provided excellent training in the development of applicable theory and algorithms. It has reawakened my inner scientist, and shaped my scientific development.

I am also grateful to Martin Wainwright and Bin Yu for their rigorous training during the initial years of my doctoral studies. In particular, I would like to thank Martin for teaching me to think and write clearly and precisely and to focus on my weaknesses as well as my strengths; and I am grateful to Bin for her lessons on the human aspect of scientific collaboration. I was also fortunate to have the opportunity to work with Aditya Guntuboyina and Sujayam Saha on inequalities between f-divergences (and play some fine games of racket cricket :-).

I owe a special thanks to the following people for helpful discussions about the work contributing to this thesis: Nicholas Boyd, Pablo Parrilo, Mahdi Soltanolkotabi, Bernd Sturmfels, and Gongguo Tang for many helpful discussions about the work in Chapter 2; Stephen Boyd and Elina Robeva for many useful conversations about the work in Chapter 3; and Sivaraman Balakrishnan, Stephen Boyd, Johannes Lederer, Elina Robeva, and Siqi Wu for many helpful discussions about the work in Chapter 4. Moreover, I have the NSF and Ben to thank for support during the course of my doctoral studies.

I have been fortunate to have a sequence of fantastic roommates (Vivek, Francois, Nick, and Elina), who have shared and enhanced my scientific curiosity. Last but certainly not least, I am eternally grateful to my family for all their love and support. I would not have made it this far without you!

Chapter 1

Introduction

What can we learn by observing nature? How can we understand and predict natural phenomena? Progress in the natural and mathematical sciences is typically made through a synergistic collaboration between experimental efforts to generate new data and theoretical efforts to distill knowledge from measurements. In this thesis we analyze the measurement process itself. In particular, we develop mathematical theory to answer the experimentalist's question:

What was the input to our measurement apparatus that generated this output?

We analyze the statistical difficulty of this inverse problem and solve specific instances with provable guarantees.

Our starting point is an information theoretic prior of parsimony: we assume the input signal is simple, with low information content. We do acknowledge that the true nature of the measured phenomenon may not even be finitely describable. It may be infinitely complex! However, since the output is only recorded with finite precision, the best we can do is produce more complex descriptions of the measured signal as we have more data available. This is why we search for a simple signal that matches the output of our measurement apparatus.

We impose simplicity on the input by assuming that it can be described by a small number of sources, and the overall measurement is the superposition of the measurement of each source. Hence, the measured signal has a sparse additive decomposition of the form

$$f = \sum_{i=1}^{M} w_i f_i.$$

Here each f_i describes the measurement of a single source, w_i is a real number (typically positive) that weighs the contribution of the ith source, and M is the number of sources. The term *sparse* refers to the fact that M is small.

We study two different measurement models in which the f_i take distinct mathematical forms. The first setup is superresolution. Here, the input is a collection of continuously parameterized sources, and we observe a weighted superposition of signals from all of the sources. The second setup is unsupervised classification. The input is a collection of categories, and the output is an unlabeled set of objects from the different categories. To build intuition, we introduce each measurement model with an example.

1. In the first setup the f_i are *functions*, and we observe the value $f(s)$ for s ranging over some set $\mathcal{S} = \{s_1, \dots, s_n\}$. For example, in superresolution fluorescence microscopy the measurement apparatus is a microscope, the input signal is a collection of fluorophores (i.e. point sources of light), and we observe the values of n pixels $f(s_1), \dots, f(s_n)$. The optics of the microscope introduce a blur around each point source, characterized by the *point spread function* ψ. The f_i are translates of ψ:

$$f_i(s) = \psi(s - t_i) \qquad \text{for } i = 1, \dots, M \quad \text{and} \quad s \in \mathcal{S}. \tag{1.1}$$

   Physically, $w\psi(s - t)$ represents the average number of photons incident on pixel s when the illumination comes from a point source of light with position t, and intensity w. The functional form of ψ is assumed known – in principle it can be derived from Maxwell's equations. The goal of superresolution microscopy is to recover the positions $t_1, \dots, t_M$, intensities $w_1, \dots, w_M$, and number M of a collection of point sources from the image

$$\left\{ \sum_{i=1}^{M} w_i \psi(s - t_i) \middle| s \in \mathcal{S} \right\}.$$

2. In the second setup the f_i are *distributions*, and we observe *samples* from the mixture distribution with mixture weights $w_i > 0$. A sample X is generated from the mixture by first drawing a label $Z \sim \text{Categorical}(w_1, \dots, w_M)$ and then generating $X \sim f_Z$. The goal is to recover the labels $Z_1, \dots, Z_n$ from the unlabeled samples $X_1, \dots, X_n$. For example, in single cell transcriptomics the measurement apparatus is a sequencer, the input signal is a population of cells with M types, and the output is a collection of gene expression profiles $X_1, \dots, X_n$. The *gene expression* of a cell is a vector X in $\mathbb{R}^d$, where d denotes the total number of genes. The entries of X denote the number of copies of mRNA for different genes. Suppose we have a population of cells of M types and with abundances $w_1, \dots, w_M$, and suppose that the gene expression of a cell of type i is drawn randomly from some distribution f_i on gene expression space. Together with the mixture weights w_i, the distributions f_i form a mixture model for gene expression profiles. The mixture model is nonparametric because we do not assume the distributions f_i come from any particular parametric family. The challenge of single cell transcriptomics is to cluster the cells by type, without prior knowledge of the distributions f_i.

There are two major differences between these setups. First, the f_i come from a *parametric* family in superresolution, but our analysis of unsupervised classification treats the f_i as nonparametric. Second, the number M of sources is unknown in superresolution, but it *is* known in our analysis of unsupervised classification. Hence we encounter distinct difficulties in our treatment of these two different setups.

In the remainder of this chapter we introduce the main results of the subsequent chapters. Section 1.1 introduces Chapters 2 and 3 which focus on continuous compressed sensing problems like superresolution microscopy. Section 1.2 introduces Chapter 4 which analyzes spectral clustering under nonparametric mixture models.

1.1 Superresolution

The emerging field of *superresolution* has applications in a wide array of empirical sciences including fluorescence microscopy, astronomy, lidar, X-ray diffraction, and electron microscopy. For example, superresolution techniques in fluorescence microscopy are revolutionizing biology by enabling direct visualization of single molecules in living cells. Superresolution is made possible by signal processing techniques that leverage *sparsity*: the assumption that an observed signal is the noisy measurement of a few weighted sources. The past decade witnessed a lot of excitement about *compressed sensing* methods that recover sparse vectors from noisy, incomplete measurements. However, almost all superresolution problems in the natural sciences involve *continuously* parameterized sources where the set of candidate parameter values is infinite. For example, the image of a collection of point sources of light is parameterized by the source locations which can vary continuously in the image plane.

The focus of Chapters 2 and 3 is on compressed sensing problems with continuous dictionaries. We develop the mathematical theory of superresolution viewed an optimization problem over the infinite dimensional space of measures. Specifically, in Chapter 2 we prove that a weighted total variation minimization scheme can recover the true source locations in ideal settings, and in Chapter 3 we develop an algorithm to solve the semi-infinite convex optimization problems that arise from this measure-theoretic formulation of superresolution. Before introducing the specific contributions, we set up the mathematical framework of superresolution.

Measurement Model

We assume the existence of an underlying set of objects called sources and a parameter space Θ. Each source has a parameter $t \in \Theta$ and a nonnegative weight $w > 0$. The signal generated by a collection of sources $\{(w_i, t_i)\}_{i=1}^M$ is given by

$$y = \sum_{i=1}^{M} w_i \phi(t_i) + \nu. \tag{1.2}$$

Here $\phi : \Theta \to \mathbb{R}^n$ is a function that describes the contribution of a single source to the measurement, and $\nu \in \mathbb{R}^n$ is an additive noise term. Our goal is to find the signal parameters $\{(w_i, t_i)\}_{i=1}^M$, and their number M, from the noisy measurement y.

To solidify intuition, we briefly illustrate the physical interpretation of this measurement model for the example of fluorescence microscopy. Suppose we take a picture of a collection of fluorescent proteins through a microscope. Each protein is essentially a point source of light, but because light diffracts as it passes through the aperture, the image is convolved with the point spread function of the microscope. The image of a fluorophore at position $t \in \mathbb{R}^2$ and with brightness $w > 0$ is

$$w\phi(t) = w \begin{bmatrix} \psi(t, s_1) \\ \vdots \\ \psi(t, s_n) \end{bmatrix},$$

where the function ψ is the point spread function of the microscope from Equation (1.1) and $\phi : \mathbb{R}^2 \to \mathbb{R}^n$ is the pixelated point spread function. The total image is the linear superposition of the images of the individual fluorophores, plus additive noise. Our goal is to remove the effects of diffraction and pixelization and recover the point source locations t_i and intensities w_i. In Section 3.2 of Chapter 3 we outline more examples.

Approach

One possible approach to recover the signal parameters $\{(w_i, t_i)\}_{i=1}^M$ is to grid the parameter space Θ and restrict attention to a finite set of candidate signal parameters. In particular, we could select grid points $g_1, \ldots, g_G \in \Theta$ and assume $t_i \in \{g_1, \ldots, g_G\}$. We can therefore write $y = \sum_{j=1}^G \omega_j \phi(g_j) + \nu$, where $\omega_j = 0$ except for M values of j where $\omega_j = w_i$ for some i. Hence the problem of recovering the signal parameters is reduced to identifying a sparse vector $\omega \in \mathbb{R}^G$. The standard approach to identify such a sparse vector is the ℓ_1 regularized sparse regression problem

$$\begin{array}{ll} \text{minimize} & \left\| \sum_{j=1}^{G} \omega_j \phi(g_j) - y \right\|_2 \\ \text{subject to} & \sum_{j=1}^{G} |\omega_j| \leq \tau. \end{array}$$

Here τ is a regularization term that controls the sparsity level of ω. However, this approach has significant drawbacks. The theoretical requirements imposed by the classical models of compressed sensing become more stringent as the grid becomes finer. Furthermore, making the grid finer can also lead to numerical instabilities and computational bottlenecks in practice.

Another potential approach is to make a guess for M, and attempt to solve

$$\begin{array}{ll} \underset{w_i, t_i}{\text{minimize}} & \left\| \sum_{i=1}^{M} w_i \phi(t_i) - y \right\|_2 \\ \text{subject to} & \sum_{i=1}^{M} |w_i| \leq \tau. \end{array}$$

However, this problem is nonconvex in (w_i, t_i) and it is not clear how to choose M. Hence, it is difficult to give theoretical guarantees, and in practice an algorithm to solve this optimization problem can be trapped in local minima.

This motivates the following measure theoretic formulation of the superresolution problem. We encode the signal parameters in a sum of Diracs $\mu_\star = \sum_{i=1}^M w_i \delta_{t_i}$, where δ_t denotes the Dirac distribution centered at t. In terms of $\mu_\star$, the measurement is $y = \int \phi(t) d\mu_\star(t) + \nu$. Our goal is to invert this operation to recover the measure $\mu_\star$ and hence recover the signal parameters encoded in $\mu_\star$.

After observing $y \in \mathbb{R}^n$, we estimate the signal parameters encoded in $\mu_\star$ by minimizing a convex loss ℓ of the residual between y and $\int \phi(t)d\mu(t)$:

$$\begin{array}{ll} \text{minimize}_\mu & \left\| y - \int \phi(t)d\mu(t) \right\|_2 \\ \text{subject to} & \|\mu\|_{\text{TV}} \leq \tau. \end{array} \tag{1.3}$$

This is a convex optimization problem over the infinite dimensional space of measures. Here $\|\mu\|_{\text{TV}}$ denotes the total variation of the measure μ, an infinite dimensional analogue of the standard convex heuristic in sparse recovery and compressed sensing problems [35].

Chapters 2 and 3 of this thesis contain two contributions to "compressed sensing off the grid". We introduce these contributions below.

Superresolution without Separation

Much of the mathematical analysis on recovery has relied heavily on the assumption that the sources are separated by at least some minimum amount, even in the absence of noise. In Chapter 2 we prove that in one dimension ($\Theta = \mathbb{R}$) and in the absence of noise, the positions of positively weighted sources can be recovered by solving a semi-infinite linear program, no matter how close they are. In the absence of noise, it is possible to achieve $y = \int \phi(t)d\mu(t)$ in the objective of (1.3). Therefore we reformulate (1.3) with $y = \int \phi(t)d\mu(t)$ as a constraint, and for the objective we minimize the weighted total variation. In particular, we prove that the solution to the following semi-infinite linear program recovers the signal parameters:

$$\begin{array}{ll} \underset{\mu \geq 0}{\text{minimize}} & \int h(t)d\mu(t) \\ \text{subject to} & y = \int \phi(t)d\mu(t). \end{array}$$

Here h is a weighting function that weights the measure at different locations.

Our proof improves on the technique of Candès and Fernandes-Granada [27], who construct a certificate of optimality by solving a certain system of linear equations. They prove that the system has a unique solution because the matrix for the system is close to the identity when the sources are well separated. The key new idea of our approach, by contrast, is to impose a *Tchebycheff system* condition to guarantee invertibility directly. Indeed, a matrix need not be close to the identity to be invertible! Another key difference is that we consider the weighed objective $\int h(t)d\mu(t)$, while prior work [27, 111] has analyzed the unweighted objective $\int d\mu(t)$. We, too, could not remove the separation condition without reweighing by $h(t)$. In Chapter 2 we provide evidence that this mathematically-motivated reweighing step actually improves performance in practice.

The Alternating Descent Conditional Gradient Method

Chapter 3 introduces a general approach to solve the infinite dimensional optimization problems that arise from our measure-theoretic formulation of superresolution. The algorithm, the alternating descent conditional gradient method (ADCG), enjoys the rapid local convergence and modeling flexibility of nonconvex programming algorithms, but also the stability and global convergence guarantees associated with convex optimization. ADCG is a measure-theoretic formulation of the CoGENT algorithm developed by Wright and Shah [90] that can leverage structure in the parameter space, and differentiable measurement models. ADCG interleaves conditional gradient steps on the convex objective with nonconvex improvement of the signal parameters and weights. For the nonconvex descent step we propose alternating descent over the weights and parameters. We show that the conditional gradient steps update μ by adding a single element to its support, and the total support remains bounded as the algorithm runs. Moreover we prove that ADCG converges, and achieves accuracy ϵ in $\mathcal{O}(1/\epsilon)$ steps. We find that the nonconvex step is the key to ADCG's good performance in practice. Indeed, without the nonconvex step, the algorithm can only change the support of μ by adding and removing points. As our theoretical analysis only leverages the fact that the nonconvex step does not worsen the result, we suspect that the convergence rate is far from tight and can be significantly improved by a deeper analysis that charaterizes the impact of the nonconvex step.

1.2 Spectral clustering

Clustering algorithms are valuable in many data driven scientific endeavors for their ability to automatically detect interpretable heterogeneity. The most basic clustering algorithms search for clusters of a particular parametric form: for example, a mixture of Gaussians. Spectral clustering, on the other hand, is a more versatile algorithm. As we shall see in Chapter 4, spectral clustering leverages a powerful preprocessing step that makes clusters easy to detect and separate.

Measurement Model

We now introduce the nonparametric mixture model formalism for unsupervised classification. Let $\mathbb{P}_1, \dots, \mathbb{P}_M$ be a collection of probability distributions on a compact set $\mathcal{X}$ and let $\{w_i\}_{i=1}^M \subset \mathbb{R}_+$ be a convex combination. That is, $\sum_{i=1}^M w_i = 1$, and $w_i > 0$. We are given n independent observations $X_1, \dots, X_n$ of the random variable distributed according to the mixture $\bar{\mathbb{P}} = \sum_{i=1}^M w_i \mathbb{P}_i$. This mixture distribution is *nonparametric* because the mixture components are not restricted to any parametric family. A random sample $X \sim \bar{\mathbb{P}}$ can be obtained by first drawing a label $Z \sim \text{Categorical}(w_1, \dots, w_M)$, and conditioned on the event $\{Z = m\}$, drawing $X \sim \mathbb{P}_m$. The unsupervised classification problem can be formalized as recovering these *latent labels* $\{Z_j\}_{j=1}^n$ from the unlabeled samples $\{X_j\}_{j=1}^n$.

To solidify intuition, it might help to keep in mind the following two natural examples: cells and stars both have *types*. First, imagine a population of cells with M types of abundance $w_1, \dots, w_M$. A randomly selected cell will have type m with probability w_m. Cells are classified by the genes they express (every cell in an organism has essentially the same DNA, but cells of different types use the DNA in different ways by expressing different genes). The mixture model postulates that the gene expression profile of a randomly selected cell of type

m as a draw from a distribution $\mathbb{P}_m$. Hence, a randomly selected cell from the population will have a gene expression profile drawn according to $\bar{\mathbb{P}} = \sum_{i=1}^M w_i \mathbb{P}_i$.

Second, imagine looking through a telescope at a region of the sky with M star types. The different star types are composed of different elements, and hence give off light of different wavelengths–their *spectra* are different in the optical sense. If the star types have abundances $w_1, \ldots, w_M$, and the emission spectrum of a star of type i is drawn from $\mathbb{P}_i$, then the emission spectrum of a randomly selected star has distribution $\sum_{i=1}^M w_i \mathbb{P}_i$.

Approach

How can we separate samples from a nonparametric mixture? Simple clustering algorithms like K-means don't work with general cluster shapes. Spectral clustering, on the other hand, leverages a powerful nonlinear transformation that tends to make clusters linearly separable.

In its modern and most popular form, the spectral clustering algorithm [83, 103] involves two steps: first, the eigenvectors of the normalized Laplacian are used to embed the dataset, and second, the K-means clustering algorithm is applied to the embedded dataset. The *normalized Laplacian* $L \in \mathbb{R}^{n \times n}$ is defined in terms of a symmetric, continuous kernel function $k : \mathcal{X} \times \mathcal{X} \to (0, \infty)$. The kernel function gives a notion of similarity between elements of $\mathcal{X}$. A canonical example is the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2)$; it is close to 1 for vectors x and x' that are relatively close, and decays to zero for pairs that are far apart. The *kernel matrix* $A \in \mathbb{R}^{n \times n}$ is the matrix of pairwise similarities $A_{ij} = k(X_i, X_j)/n$. The normalized Laplacian L is obtained from A by a similarity transformation

$$L = D^{-1/2} A D^{-1/2},$$

where D is the diagonal matrix of row sums of A

$$D_{ii} = \frac{1}{n} \sum_{j=1}^{n} k(X_i, X_j).$$

Spectral clustering transforms the data according to the map

$$X_i \mapsto r_i,$$

where r_i is the ith row of the matrix $V = \begin{bmatrix} v_1 \ldots v_M \end{bmatrix}$, and $v_1, \ldots, v_M$ are the principal eigenvectors of L corresponding to the largest M eigenvalues. The second step of spectral clustering applies K-means to the transformed dataset $r_1, \ldots, r_n$.

The Geometry of Kernelized Spectral Clustering

To gain some intuition for why the eigenvectors of L contain information about the mixture components $\mathbb{P}_i$, note that the top eigenvector of L is

$$Lq = q,$$

where $q \in \mathbb{R}^n$ has entries $q(X_j) = \sqrt{\frac{1}{n}\sum_{i=1}^n k(X_i, X_j)}$, for $j = 1, \dots, n$. Moreover, a similar story holds for the subset of the data $S_m = \{X_i : Z_i = m\}$. If we denote the normalized Laplacian constructed from the subset S_m by L_m then we have

$$L_m q_m = q_m,$$

where $q_m \in \mathbb{R}^{|S_m|}$ has entries $q_m(X_j) = \sqrt{\frac{1}{n}\sum_{X \in S_m} k(X, X_j)}$ for $X_j \in S_m$.

Suppose for the moment that the components $\mathbb{P}_i$ have *zero* overlap with respect to the kernel k (in the sense that $k(X_i, X_j) = 0$ for all i, j such that $Z_i \neq Z_j$). Then the kernel matrix A is block diagonal under some permutation of its rows and columns, and the Laplacian L is block diagonal with blocks $L_1, \dots, L_M$. If we extend the vectors q_m to $\mathbb{R}^n$ by defining $q_m(X_i) = 0$ for $i \notin S_m$, then q_m are all top eigenvectors of L:

$$Lq_m = q_m \qquad \text{for } m = 1, \dots, M.$$

In this ideal situation of zero overlap, the top M eigenvectors of L reveal the latent labels $Z_1, \dots, Z_n$ since

$$\{j : q_m(X_j) \neq 0\} = \{j : Z_j = m\} \qquad \text{for } m = 1, \dots, M.$$

Moreover,

$$r_i = [q_1(X_i), \dots, q_M(X_i)] \qquad \text{for } m = 1, \dots, M, \tag{1.4}$$

and the embedded image of points with distinct labels are orthogonal!

Chapter 4 examines the more realistic setting where the mixture components overlap. The analysis of Chapter 4 establishes that an approximate version of the relationship (1.4) holds as long as the mixture components don't overlap too much. Chapter 4 begins by providing a novel and useful characterization of the principal eigenspace of the population-level normalized Laplacian operator: more precisely, when the mixture components are indivisible and have small overlap, the eigenspace is close to the span of the square root kernelized component densities. We then use this characterization to analyze the geometric structure of the embedding of a finite set of i.i.d. samples. Our main result is to establish a certain geometric property of nonparametric mixtures referred to as *orthogonal cone structure*. In particular, we show that when the mixture components are indivisible and have small overlap, embedded samples from different components are almost orthogonal with high probability. We then prove that this geometric structure allows K-means to correctly label most of the samples. Our proofs rely on techniques from operator perturbation theory, empirical process theory, and spectral graph theory.

Chapter 2

Superresolution without Separation

This chapter provides a theoretical analysis of diffraction-limited superresolution, demonstrating that arbitrarily close point sources can be resolved in ideal situations. Precisely, we assume that the incoming signal is a linear combination of M shifted copies of a known waveform with unknown shifts and amplitudes, and one only observes a finite collection of evaluations of this signal. We characterize properties of the base waveform such that the exact translations and amplitudes can be recovered from $2M+1$ observations. This recovery can be achieved by solving a weighted version of basis pursuit over a continuous dictionary. Our analysis shows that ℓ_1-based methods enjoy the same separation-free recovery guarantees as polynomial root finding techniques such as Prony's method or Vetterli's method for signals of finite rate of innovation. Our proof techniques combine classical polynomial interpolation techniques with contemporary tools from compressed sensing.

This chapter is joint work with Elina Robeva and Benjamin Recht. The content of this chapter has been submitted for publication under the title *Superresolution without Separation* and is available on the arxiv http://arxiv.org/abs/1506.03144.

2.1 Introduction

Imaging below the diffraction limit remains one of the most practically important yet theoretically challenging problems in signal processing. Recent advances in superresolution imaging techniques have made substantial progress towards overcoming these limits in practice [46, 84], but theoretical analysis of these powerful methods remains elusive. Building on polynomial interpolation techniques and tools from compressed sensing, this chapter provides a theoretical analysis of diffraction-limited superresolution, demonstrating that arbitrarily close point sources can be resolved in ideal situations.

We assume that the measured signal takes the form

$$x(s) = \sum_{i=1}^{M} w_i \psi(s, t_i), \tag{2.1.1}$$

Here $\psi(s,t)$ is a differentiable function that describes the image at spatial location s of a point source of light localized at t. The function ψ is called the *point spread function*, and we assume its particular form is known beforehand. In (2.1.1), $t_1, \ldots, t_M$ are the locations

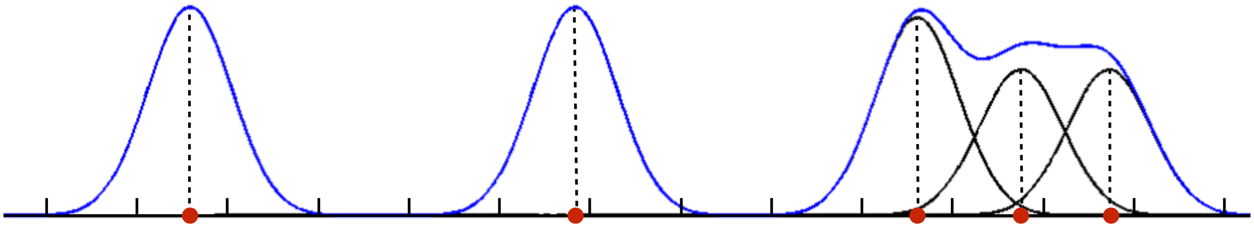

Figure 2.1: An illustrative example of (2.1.1) with the Gaussian point spread function $\psi(s,t) = e^{-(s-t)^2}$. The t_i are denoted by red dots, and the true intensities w_i are illustrated by vertical, dashedd black lines. The super position resulting in the signal x is plotted in blue. The samples $\mathcal{S}$ would be observed at the tick marks on the horizontal axis.

of the point sources and $w_1, ..., w_M > 0$ are their intensities. Throughout we assume that these quantities together with the number of point sources M, are fixed but unknown. The primary goal of superresolution is to recover the locations and intensities from a set of noiseless observations

$$\{x(s) \mid s \in \mathcal{S}\}\,.$$

Here $\mathcal{S}$ is the set of points at which we observe x; we denote the elements of $\mathcal{S}$ by $s_1, \ldots, s_n$. A mock-up of such a signal x is displayed in Figure 2.1.

In this chapter, building on the work of Candès and Fernadez-Granda [26, 27, 48] and Tang *et al* [19, 111, 113], we aim to show that we can recover the tuple (t_i, w_i, M) by solving a convex optimization problem. We formulate the superresolution imaging problem as an infinite dimensional optimization over measures. Precisely, note that the observed signal can be rewritten as

$$x(s) = \sum_{i=1}^{M} w_i \psi(s, t_i) = \int \psi(s,t) d\mu_\star(t)\,. \tag{2.1.2}$$

Here, $\mu_\star$ is the positive discrete measure $\sum_{i=1}^{M} w_i \delta_{t_i}$, where δ_t denotes the Dirac measure centered at t. We aim to show that we can recover $\mu_\star$ by solving the following:

$$\begin{aligned} \underset{\mu}{\text{minimize}} \quad & \int h(t) d\mu(t) \\ \text{subject to} \quad & x(s) = \int \psi(s,t) d\mu(t), \quad s \in \mathcal{S} \\ & \operatorname{supp} \mu \subset B \\ & \mu \geq 0\,. \end{aligned} \tag{2.1.3}$$

Here, B is a fixed compact set and $h(t)$ is a weighting function that weights the measure at different locations. The optimization problem (2.1.3) is over the set of all positive finite measures μ supported on B.

The optimization problem (2.1.2) is an analog of weighted ℓ_1 minimization over the continuous domain B. Indeed, if we know a priori that the t_i are elements of a finite discrete set Ω, then optimizing over all measures subject to $\operatorname{supp} \mu \subset \Omega$ is precisely equivalent to weighted ℓ_1 minimization. This infinite dimensional analog with uniform weights has proven useful for compressed sensing over continuous domains [113], resolving diffraction-limited images from low-pass signals [26, 48, 111], system identification [101], and many other applications [35]. We will see below that the weighting function essentially ensures that all of the candidate locations are given equal influence in the optimization problem.

Our main result, Theorem 4.3.2, establishes that for one-dimensional signals, under rather mild conditions, we can recover $\mu_\star$ from the optimal solution of (2.1.3). Our conditions, described in full-detail below, essentially require the observation of at least $2M+1$ samples, and that the set of translates of the point spread function forms a linearly independent set. In Theorem 2.1.1 we verify that these conditions are satisfied by the Gaussian point spread function for any M source locations with no minimum separation condition. This is the first analysis of an ℓ_1 based method that matches the separation-free performance of polynomial root finding techniques [117, 41, 87]. Our motivation for such an analysis is that ℓ_1 based methods generalize to higher dimensions and are empirically stable in the presence of noise.

In Chapter 3 we show that the problem (2.1.3) can be optimized to precision ϵ in polynomial time using a greedy algorithm. In our experiments in Section 2.3, we use this algorithm to demonstrate that our theory applies, and show that even in multiple dimensions with noise, we can recover closely spaced point sources.

Main Result

We restrict our theoretical attention in this Chapter to the one-dimensional case, leaving the higher-dimensional cases to future work. Let $\psi : \mathbb{R}^2 \to \mathbb{R}$ be our one dimensional point spread function, with the first argument denoting the position where we are observing the image of a point source located at the second argument. We assume that ψ is differentiable in both arguments.

For convenience, we will assume that $B = [-T, T]$ for some large scalar T. However, our proof will trivially extend to more restricted subsets of the real line. Moreover, we will state our results for the special case where $\mathcal{S} = \{s_1, \ldots, s_n\}$, although our proof is written for possibly infinite measurement sets. We define the weighting function in the objective of our optimization problem via

$$h(t) = \frac{1}{n} \sum_{i=1}^{n} \psi(s_i, t)\,.$$

Our main result establishes conditions on ψ such that the true measure $\mu_\star$ is the unique optimal solution of (2.1.3). Importantly, we show that these conditions are satisfied by the Gaussian point spread function with no separation condition.

Theorem 2.1.1. *Suppose $|\mathcal{S}| > 2M$, and $\psi(s,t) = e^{-(s-t)^2}$. Then for any $t_1 < \ldots < t_M$, the true measure $\mu_\star$ is the unique optimal solution of* (2.1.3).

Before we proceed to state the main result, we need to introduce a bit of notation and define the notion of a Tchebycheff system. Let $K(t,\tau) = \frac{1}{n}\sum_{i=1}^{n} \psi(s_i,t)\psi(s_i,\tau)$, and define the vector valued function $v : \mathbb{R} \to \mathbb{R}^{2M}$ via

$$v(s) = \begin{bmatrix} \psi(s,t_1) & \ldots & \psi(s,t_M) & \frac{d}{dt_1}\psi(s,t_1) & \ldots & \frac{d}{dt_M}\psi(s,t_M) \end{bmatrix}^T . \tag{2.1.4}$$

Definition 1. A set of functions $u_1, \ldots, u_n$ is called a *Tchebycheff system* (or *T-system*) if for any points $\tau_1 < \ldots < \tau_n$, the matrix

$$\begin{pmatrix} u_1(\tau_1) & \ldots & u_1(\tau_n) \\ \vdots & & \\ u_n(\tau_1) & \ldots & u_n(\tau_n) \end{pmatrix}$$

is invertible.

Conditions 2.1.2. *We impose the following three conditions on the point spread function ψ:*

Positivity *For all $t \in B$ we have $h(t) > 0$.*

Independence *The matrix $\frac{1}{n}\sum_{i=1}^{n} v(s_i)v(s_i)^T$ is nonsingular.*

T-system $\{K(\cdot, t_1), \ldots, K(\cdot, t_M), \frac{d}{dt_1}K(\cdot, t_1), \ldots, \frac{d}{dt_M}K(\cdot, t_M), h(\cdot)\}$ *form a T-system.*

Theorem 2.1.3. *If ψ satisfies the Conditions 2.1.2 and $|\mathcal{S}| > 2M$, then the true measure $\mu_\star$ is the unique optimal solution of (2.1.3).*

Note that the first two parts of Conditions 2.1.2 are easy to verify. Positivity eliminates the possibility that a candidate point spread function could equal zero at all locations—obviously we would not be able to recover the source in such a setting! Independence is satisfied if

$$\{\psi(\cdot, t_1), \ldots, \psi(\cdot, t_M), \frac{d}{dt_1}\psi(\cdot, t_1), \ldots, \frac{d}{dt_M}\psi(\cdot, t_M)\} \quad \text{is a T-system.}$$

This condition allows us to recover the amplitudes uniquely assuming we knew the true t_i locations *a priori*, but it is also useful for constructing a *dual certificate* as we discuss below.

We remark that we actually prove the theorem under a weaker condition than T-system. Define the matrix-valued function $\Lambda : \mathbb{R}^{2M+1} \to \mathbb{R}^{2M+1\times 2M+1}$ by

$$\Lambda(p_1, \ldots, p_{2M+1}) := \begin{bmatrix} \kappa(p_1) & \ldots & \kappa(p_{2M+1}) \\ 1 & \ldots & 1 \end{bmatrix}, \tag{2.1.5}$$

where $\kappa : \mathbb{R} \to \mathbb{R}^{2M}$ is defined as

$$\kappa(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\psi(s_i, t)}{h(t)} v(s_i)\,. \tag{2.1.6}$$

Our proof of Theorem 4.3.2 replaces condition T-system with the following:

Determinantal There exists $\rho > 0$ such that for any $t_i^-, t_i^+ \in (t_i - \rho, t_i + \rho)$, and $t \in [-T, T]$,
the matrix $\Lambda\big(t_1^-, t_1^+, \ldots, t_M^-, t_M^+, t\big)$ is nonsingular whenever t, t_i^-, t_i^+ are distinct.

This condition looks more complicated than T-SYSTEM and is indeed nontrivial to verify. It is essentially a *local T-system* condition in the sense that the points τ_i Definition 1 are restricted to lie in a small neighborhood about the t_i. It is clear that T-SYSTEM implies DETERMINANTAL. The advantage of the more general condition is that it can hold for finitely supported ψ, while this is not true for T-SYSTEM. In fact, it is easy to see that if T-SYSTEM holds for any point spread function ψ, then DETERMINANTAL holds for the truncated version $\psi(s,t)\mathbf{1}\{|s-t| \leq 3T\}$, where $\mathbf{1}\{x \leq y\}$ is the indicator variable equal to 1 when $x \leq y$ and zero otherwise. We suspect that DETERMINANTAL may hold for significantly tighter truncations.

As we will see below, T-SYSTEM and INDEPENDENCE are related to the existence of a canonical *dual certificate* that is used ubiquitously in sparse approximation [29, 51]. In compressed sensing, this construction is due to Fuchs [51], but its origins lie in the theory of polynomial interpolation developed by Markov and Tchebycheff, and extended by Gantmacher, Krein, Karlin and others (see the survey in Section 2.1).

In the continuous setting of superresolution, the dual certificate becomes a *dual polynomial*: a function of the form $Q(t) = \frac{1}{n}\sum_{j=1}^{n} \psi(s_j,t)q(s_j)$ satisfying

$$\begin{aligned} Q(t) &\leq h(t) \\ |Q(t_i)| &= h(t), \quad i = 1, \dots, M. \end{aligned} \tag{2.1.7}$$

To see how T-SYSTEM might be useful for constructing a dual polynomial, note that as $t_1^+ \downarrow t_1$ and $t_1^- \uparrow t_1$, the first two columns of $\Lambda(t_1^+, t_1^-, \dots, t)$ converge to the same column, namely $\kappa(t_1)$. However, if we divide by the difference $t_1^+ - t_1^-$, and take a limit then we obtain the derivative of the second column. In particular, some calculation shows T-SYSTEM implies

$$\det \begin{bmatrix} A & \kappa(t) \\ \omega & h(t) \end{bmatrix} \neq 0 \qquad \forall t \neq t_i,$$

where $A = \frac{1}{n}\sum_{j=1}^{n} v(s_i)v(s_i)^T$ is the matrix from INDEPENDENCE, and

$$\omega = [h(t_1), \dots, h(t_M), h'(t_1), \dots, h'(t_M)].$$

Taking the Schur complement in $h(t)$, we find

$$\det \begin{bmatrix} A & \kappa(t) \\ \omega & h(t) \end{bmatrix} = \det A \left[\omega^T A^{-1}\kappa(t) - h(t)\right].$$

Hence it seems like the function $\omega^T A^{-1}\kappa(t)$ might serve well as our dual polynomial. However, it remains unclear from this short calculation that this function is bounded above by $h(t)$. The proof of Theorem 4.3.2 makes this construction rigorous using the theory of T-systems.

Before turning to the proofs of these theorems (c.f. Sections 2.2 and 2.2), we survey the mathematical theory of superresolution imaging.

Foundations: Tchebycheff Systems

Our proofs rely on the machinery of Tchebycheff[1] systems. This line of work originated in the 1884 doctoral thesis of A. A. Markov on approximating the value of an integral $\int_a^b f(x)dx$ from the moments $\int_a^b xf(x)dx, \ldots, \int_a^b x^n f(x)dx$. His work formed the basis of the proof by Tchebycheff (who was Markov's doctoral advisor) of the central limit theorem in 1887 [114].

Recall that we defined a T-system in Definition 1. An equivalent definition of a T-system is: *the functions $u_1, ..., u_n$ form a T-system if and only if every linear combination $U(t) = a_1u_1(t) + \cdots + a_nu_n(t)$ has at most $n-1$ zeros.* One natural example of a T-system is given by the functions $1, t, \ldots, t^{n-1}$. Indeed, a polynomial of order $n-1$ can have at most $n-1$ zeros. Equivalently, the Vandermonde determinant does not vanish,

$$\begin{vmatrix} 1 & 1 & \ldots & 1 \\ t_1 & t_2 & \ldots & t_n \\ t_1^2 & t_2^2 & \ldots & t_n^2 \\ \vdots & & & \\ t_1^{n-1} & t_2^{n-1} & \ldots & t_n^{n-1} \end{vmatrix} \neq 0,$$

for any $t_1 < \ldots < t_n$. Just as Vandermonde systems are used to solve polynomial interpolation problems, T-systems allows the generalization of the tools from polynomial fitting to a broader class of nonlinear function-fitting problems. Indeed, given a T-system $u_1, ..., u_n$, a *generalized polynomial* is a linear combination $U(t) = a_1u_1(t) + \cdots + a_nu_n(t)$. The machinery of T-systems provides a basis for understanding the properties of these generalized polynomials. For a survey of T-systems and their applications in statistics and approximation theory, see [52, 65, 66]. In particular, many of our proofs are adapted from [66], and we call out the parallel theorems whenever this is the case.

Prior art and related work

Broadly speaking, superresolution techniques enhance the resolution of a sensing system, optical or otherwise; *resolution* is the distance at which distinct sources appear indistinguishable. The mathematical problem of localizing point sources from a blurred signal has applications in a wide array of empirical sciences: astronomers deconvolve images of stars to angular resolution beyond the Rayleigh limit [89], and biologists capture nanometer resolution images of fluorescent proteins [21, 57, 97, 72]. Detecting neural action potentials from extracellular electrode measurements is fundamental to experimental neuroscience [45], and resolving the poles of a transfer function is fundamental to system identification [101]. To understand a radar signal, one must decompose it into reflections from different sources [55]; and to understand an NMR spectrum, one must decompose it into signatures from different chemicals [111].

The mathematical analysis of point source recovery has a long history going back to the work of Prony [87] who pioneered techniques for estimating sinusoidal frequencies. Prony's

[1] Tchebycheff is one among many transliterations from the cyrillic. Others include Chebyshev, Chebychev, and Cebysev.

method is based on algebraically solving for the roots of polynomials, and can recover arbitrarily closely spaced frequencies. The annihilation filter technique introduced by Vetterli [117] can perfectly recover any signal of *finite rate of innovation* with minimal samples. In particular the theory of signals with finite rate of innovation shows that given a superposition of pulses of the form $\sum a_k \psi(t-t_k)$, one can reconstruct the shifts t_k and coefficients a_k from a minimal number of samples [41, 117]. This holds without any separation condition on the t_k and as long as the base function ψ can reproduce polynomials of a certain degree (see [41, Section A.1] for more details). The algorithm used for this reconstruction is however based on polynomial rooting techniques that do not easily extend to higher dimensions. Moreover, this algebraic technique is not robust to noise (see the discussion in [110, Section IV.A] for example).

In contrast we study sparse recovery techniques. This line of thought goes back at least to Carathéodory [33, 32]. Our contribution is an analysis of ℓ_1 based methods that matches the performance of the algebraic techniques of Vetterli in the one dimensional and noiseless setting. Our primary motivation is that ℓ_1 based methods may be more stable to noise and trivially generalize to higher dimensions (although our analysis currently does not).

It is tempting to apply the theory of compressed sensing [10, 29, 30, 38] to problem (2.1.3). If one assumes the point sources are located on a finite grid and are well separated, then some of the standard models for recovery are valid (e.g. incoherency, restricted isometry property, or restricted eigenvalue property). With this motivation, many authors solve the gridded form of the superresolution problem in practice [9, 11, 78, 42, 47, 56, 91, 108, 109, 72, 43]. However, this approach has some significant drawbacks. The theoretical requirements imposed by the classical models of compressed sensing become more stringent as the grid becomes finer. Furthermore, making the grid finer can also lead to numerical instabilities and computational bottlenecks in practice.

Despite recent successes in many empirical disciplines, the theory of superresolution imaging remains limited. Candès and Fernandes-Granada [27] recently made an important contribution to the mathematical analysis of superresolution, demonstrating that semi-infinite optimization could be used to solve the classical Prony problem. Their proof technique has formed the basis of several other analyses including that of Bendory *et al* [17] and that of Tang *et al* [111]. To better compare with our approach, we briefly describe the approach of [17, 27, 111] here.

They construct the vector q of a dual polynomial $Q(t) = \frac{1}{n}\sum_{j=1}^{n} \psi(s_j, t) q_j$ as a linear combination of $\psi(s, t_i)$ and $\frac{d}{dt_i}\psi(s, t_i)$. In particular, they define the coefficients of this linear combination as the least squares solution to the system of equations

$$\begin{aligned} Q(t_i) &= \operatorname{sign}(w_i), && i = 1, \ldots, M \\ \frac{d}{dt}Q(t)\Big|_{t=t_i} &= 0, && i = 1, \ldots, M. \end{aligned} \tag{2.1.8}$$

They prove that, under a minimum separation condition on the t_i, the system has a unique solution because the matrix for the system is a perturbation of the identity, hence invertible.

Much of the mathematical analysis on superresolution has relied heavily on the assumption that the point sources are separated by more than some minimum amount [14, 17, 27, 40, 44, 81, 39]. We note that in practical situations with noisy observations, some form of minimum separation may be necessary. One can expect, however, that the required minimum separation should go to zero as the noise level decreases: a property that is not manifest in previous results. Our approach, by contrast, does away with the minimum separation condition by observing that the matrix for the system (2.1.8) need not be close to the identity to be invertible. Instead, we impose Conditions 2.1.2 to guarantee invertibility directly. Not surprisingly, we use techniques from T-systems to construct an analog of the polynomial Q in (2.1.8) for our specific problem.

Another key difference is that we consider the weighted objective $\int h(t)d\mu(t)$, while prior work [17, 27, 111] has analyzed the unweighted objective $\int d\mu(t)$. We, too, could not remove the separation condition without reweighing by $h(t)$. In Section 2.3 we provide evidence that this mathematically motivated reweighing step actually improves performance in practice. Weighting has proven to be a powerful tool in compressed sensing, and many works have shown that weighting an ℓ_1-like cost function can yield improved performance over standard ℓ_1 minimization [50, 68, 116, 22]. To our knowledge, the closest analogy to our use of weights comes from Rauhut and Ward, who use weights to balance the influence of dynamic range of bases in polynomial interpolation problems [92]. In the setting of this chapter, weights will serve to lessen the influence of sources that have low overlap with the observed samples.

We are not the first to bring the theory of Tchebycheff systems to bear on the problem of recovering finitely supported measures. De Castro and Gamboa [34] prove that a finitely supported positive measure μ can be recovered exactly from measurements of the form

$$\left\{ \int u_0 d\mu, \ldots, \int u_n d\mu \right\}$$

whenever $\{u_0, \ldots, u_n\}$ form a T-system containing the constant function $u_0 = 1$. These measurements are almost identical to ours; if we set $u_k(t) = \psi(s_k, t)$ for $k = 1, \ldots, n$, where $\{s_1, \ldots, s_n\} = \mathcal{S}$ is our measurement set, then our measurements are of the form

$$\{x(s) \mid s \in \mathcal{S}\} = \left\{ \int u_1 d\mu, \ldots, \int u_n d\mu \right\}.$$

However, in practice it is often impossible to directly measure the mass $\int u_0 d\mu = \int d\mu$ as required by (2.1). Moreover, the requirement that $\{1, \psi(s_1, t), \ldots, \psi(s_n, t)\}$ form a T-system does not hold for the Gaussian point spread function $\psi(s, t) = e^{-(s-t)^2}$ (see Remark 2.2). Therefore the theory of [34] is not readily applicable to superresolution imaging.

We conclude our review of the literature by discussing some prior literature on ℓ_1-based superresolution without a minimum separation condition. We would like to mention the work of Fuchs [51] in the case that the point spread function is band-limited and the samples are on a regularly-spaced grid. This result also does not require a minimum separation condition. However, our results hold for considerably more general point spread functions and sampling patterns. Finally, in a recent paper Bendory [16] presents an analysis of ℓ_1 minimization in a discrete setup by imposing a Rayleigh regularity condition which, in the absence of noise, requires no minimum separation. Our results are of a different flavor, as our setup is continuous. Furthermore we require linear sample complexity while the theory of Bendory [16] requires infinitely many samples.

2.2 Proofs

In this section we prove Theorem 4.3.2 and Theorem 2.1.1. We start by giving a short list of notation to be used throughout the proofs. We write our proofs for an arbitrary measurement $\mathcal{S}$ which need not be finite for the sake of the proof. Let P denote a fixed positive measure on $\mathcal{S}$, and set

$$h(t) = \int \psi(s,t)dP(s).$$

For concreteness, the reader might think of P as the uniform measure over $\mathcal{S}$, where if $\mathcal{S}$ is finite then $h(t) = \frac{1}{n}\sum_{j=1}^{n} \psi(s_j, t)$. Just note that the particular choice of P does not affect the proof.

Notation Glossary

- We denote the inner product of functions $f, g \in L_P^2$ by $\langle f, g\rangle_P := \int f(s)g(s)dP(s)$.
- For any differentiable function $f : \mathbb{R}^2 \to \mathbb{R}$, we denote the derivative in its first argument by $\partial_1 f$ and in its second argument by $\partial_2 f$.
- For $t \in \mathbb{R}$, let $\psi_t(\cdot) = \psi(\cdot, t)$.

Proof of Theorem 4.3.2

We prove Theorem 4.3.2 in two steps. We first reduce the proof to constructing a function q such that $\langle q, \psi_t\rangle_P$ possesses some specific properties.

Proposition 2.2.1. *If the first three items of Conditions 2.1.2 hold, and if there exists a function q such that $Q(t) := \langle q, \psi_t\rangle_P$ satisfies*

$$\begin{aligned} Q(t_j) &= h(t_j), \quad j = 1, \dots, M \\ Q(t) &< h(t_j), \quad \textit{for } t \in [-T, T] \textit{ and } t \neq t_j, \end{aligned} \tag{2.2.1}$$

then the true measure $\mu_\star := \sum_{j=1}^{M} c_j\delta_{t_j}$ is the unique optimal solution of the program 2.1.3.

This proof technique is somewhat standard [29, 51]: the function $Q(t)$ is called a *dual certificate* of optimality. However, introducing the function $h(t)$ is a novel aspect of our proof. The majority of arguments have $h(t) = 1$. Note that when $\int \psi(s,t)dP(s)$ is independent of t, then $h(t)$ is a constant and we recover the usual method of proof.

In the second step we construct $q(s)$ as a linear combination of the t_i-centered point spread functions $\psi(s, t_i)$ and their derivatives $\partial_2\psi(s, t_i)$.

Theorem 2.2.2. *Under the Conditions 2.1.2, there exist $\alpha_1, \dots, \alpha_M, \beta_1, \dots, \beta_M, c \in \mathbb{R}$ such that $Q(t) = \langle q, \psi_t\rangle_P$ satisfies (2.2.1), where*

$$q(s) = \sum_{i=1}^{M}(\alpha_i\psi(s,t_i) + \beta_i \frac{d}{dt_i}\psi(s,t_i)) + c.$$

To complete the proof of Theorem 4.3.2, it remains to prove Proposition 2.2.1 and Theorem 4.3.1. Their proofs can be found in Sections 2.2 and 2.2 respectively.

Proof of Proposition 2.2.1

We show that $\mu_\star$ is the optimal solution of problem (2.1.3) through strong duality. The Lagrangian for (2.1.3) is

$$L(q,\mu) = \int h(t)d\mu(t) + \int_{\mathcal{S}} q(s)\left(x(s) - \int \psi(s,t)d\mu(t)\right)dP(s),$$

where $q \in L_P^2$ is the dual variable. Hence the dual of problem (2.1.3) is

$$\underset{q}{\text{maximize}} \;\; \underset{\substack{\mu\geq 0\\ \operatorname{supp}\mu\subset[-T,T]}}{\text{minimize}} \;\; L(q,\mu).$$

The dual of problem (2.1.3) can be written as

$$\begin{aligned} &\text{maximize}_q && \langle q, x\rangle_P \\ &\text{subject to} && \langle q, \psi_t\rangle_P \leq h(t) \quad \text{for } t \in [-T,T]. \end{aligned} \tag{2.2.2}$$

Since the primal (2.1.3) is equality constrained, Slater's condition naturally holds, implying strong duality. As a consequence, we have

$$\langle q, x\rangle_P = \int h(t)d\mu(t) \iff q \text{ is dual optimal and } \mu \text{ is primal optimal.}$$

Suppose q satisfies (2.2.1). Hence q is dual feasible and we have

$$\begin{aligned} \langle q, x\rangle_P &= \sum_{j=1}^{M} w_j \langle q, \psi_{t_j}\rangle_P = \sum_{j=1}^{M} w_j Q(t_j) \\ &= \int h(t)d\mu_\star(t). \end{aligned}$$

Therefore, q is dual optimal and $\mu_\star$ is primal optimal.

Next we show uniqueness. Suppose the primal (2.1.3) has another optimal solution

$$\hat{\mu} = \sum_{j=1}^{\hat{M}} \hat{w}_j \delta_{\hat{t}_j}$$

such that $\{\hat{t}_1, \ldots, \hat{t}_{\hat{M}}\} \neq \{t_1, \ldots, t_M\} := \mathcal{T}$. Then we have

$$\begin{aligned} \langle q, x\rangle_P &= \sum_j \hat{w}_j \left\langle q, \psi_{\hat{t}_j}\right\rangle_P \\ &= \sum_{\hat{t}_j\in\mathcal{T}} \hat{w}_j Q(\hat{t}_j) + \sum_{\hat{t}_j\notin\mathcal{T}} \hat{w}_j Q(\hat{t}_j) \\ &< \sum_{\hat{t}_j\in\mathcal{T}} \hat{w}_j h(\hat{t}_j) + \sum_{\hat{t}_j\notin\mathcal{T}} \hat{w}_j h(\hat{t}_j) = \int h(t)d\hat{\mu}(t). \end{aligned}$$

Therefore, all optimal solutions must be supported on $\{t_1, \dots, t_M\}$.

We now show that the coefficients of any optimal $\hat{\mu}$ are uniquely determined. By condition INDEPENDENCE the matrix $\int v(s)v(s)^T dP(s)$ is invertible. Since it is also positive semidefinite, then it is positive definite, so, in particular its upper $M \times M$ block is also positive definite.

$$\det \int \begin{bmatrix} \psi(s,t_1) \\ \vdots \\ \psi(s,t_M) \end{bmatrix} \begin{bmatrix} \psi(s,t_1) & \dots & \psi(s,t_M) \end{bmatrix} dP(s) \neq 0.$$

Hence there must be $s_1, \dots, s_M \in \mathcal{S}$ such that the matrix with entries $\psi(s_i,t_j)$ is nonsingular.

Now consider some optimal $\hat{\mu} = \sum_{i=1}^M \hat{w}_i t_i$. Since $\hat{\mu}$ is feasible we have

$$x(s_j) = \sum_{i=1}^M \hat{w}_i \psi(s_j,t_i) = \sum_{i=1}^M w_i \psi(s_j,t_i) \quad \text{for} \quad j = 1, \dots, M.$$

Since $\psi(s_i,t_j)$ is invertible, we conclude that the coefficients $w_1, \dots, w_M$ are unique. Hence $\mu_\star$ is the unique optimal solution of (2.1.3).

Proof of Theorem 4.3.1

We construct $Q(t)$ via a limiting interpolation argument due to Krein [71]. We have adapted some of our proofs (with nontrivial modifications) from the aforementioned text by Karlin and Studden [66]. We give reference to the specific places where we borrow from classical arguments.

In the sequel, we make frequent use of the following elementary manipulation of determinants:

Lemma 2.2.3. *If $v_0, \dots, v_n$ are vectors in $\mathbb{R}^n$, and n is even, then*

$$\begin{vmatrix} v_1 - v_0 & \dots & v_n - v_0 \end{vmatrix} = \begin{vmatrix} v_1 & \dots & v_n & v_0 \\ 1 & \dots & 1 & 1 \end{vmatrix}.$$

Proof. By elementary manipulations, both determinants in the statement of the lemma are equal to

$$\begin{vmatrix} v_1 - v_0 & \dots & v_n - v_0 & v_0 \\ 0 & \dots & 0 & 1 \end{vmatrix}.$$

□

In what follows, we consider $\epsilon > 0$ such that

$$t_1 - \epsilon < t_1 + \epsilon < t_2 - \epsilon < t_2 + \epsilon < \dots < t_M - \epsilon < t_M + \epsilon.$$

Definition 2. A point t is a *nodal* zero of a continuous function $f : \mathbb{R} \to \mathbb{R}$ if $f(t) = 0$ and f changes sign at t. A point t is a *non-nodal* zero if $f(t) = 0$ but f does not change sign at t. This distinction is illustrated in Figure 2.2.

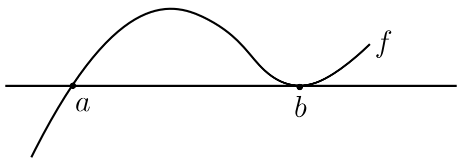

Figure 2.2: The point a is a *nodal* zero of f, and the point b is a *non-nodal* zero of f.

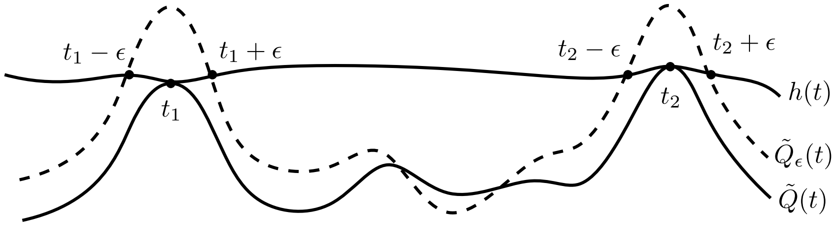

Figure 2.3: The relationship between the functions $h(t)$, $\tilde{Q}_\epsilon(t)$ and $\tilde{Q}(t)$. The function $\tilde{Q}_\epsilon(t)$ touches $h(t)$ only at $t_i \pm \epsilon$, and these are nodal zeros of $\tilde{Q}_\epsilon(t) - h(t)$. The function $\tilde{Q}(t)$ touches $h(t)$ only at t_i and these are non-nodal zeros of $\tilde{Q}(t) - h(t)$.

Our proof of Theorem 4.3.1 proceeds as follows. With ϵ fixed, we construct a function

$$\tilde{Q}_\epsilon(t) = \sum_{i=1}^{M} \alpha_\epsilon^{[i]} K_P(t, t_i) + \beta_\epsilon^{[i]} \partial_2 K_P(t, t_i)$$

such that $\tilde{Q}_\epsilon(t) = h(t)$ only at the points $t = t_j \pm \epsilon$ for all $j = 1, 2, \ldots, M$ and the points $t_j \pm \epsilon$ are nodal zeros of $\tilde{Q}_\epsilon(t) - h(t)$ for all $j = 1, 2, \ldots, M$. We then consider the limiting function $\tilde{Q}(t) = \lim_{\epsilon \downarrow 0} \tilde{Q}_\epsilon(t)$, and prove that either $\tilde{Q}(t)$ satisfies (2.2.1) or $2h(t) - \tilde{Q}(t)$ satisfies (2.2.1). An illustration of this construction is pictured in Figure 2.3.

We begin with the construction of $\tilde{Q}_\epsilon$. We aim to find the coefficients $\alpha_\epsilon, \beta_\epsilon$ to satisfy

$$\tilde{Q}_\epsilon(t_i - \epsilon) = h(t_i - \epsilon) \quad \text{and} \quad \tilde{Q}_\epsilon(t_i + \epsilon) = h(t_i + \epsilon) \quad \text{for} \quad i = 1, \ldots, M.$$

This system of equations is equivalent to the system

$$\begin{aligned} \tilde{Q}_\epsilon(t_i - \epsilon) &= h(t_i - \epsilon) \quad \text{for} \quad i = 1, \ldots, M \\ \frac{\tilde{Q}_\epsilon(t_i + \epsilon) - \tilde{Q}_\epsilon(t_i - \epsilon)}{2\epsilon} &= \frac{h(t_i + \epsilon) - h(t_i - \epsilon)}{2\epsilon} \quad \text{for} \quad i = 1, \ldots, M. \end{aligned} \tag{2.2.3}$$

Note that this is a linear system of equations in $\alpha_\epsilon, \beta_\epsilon$ with coefficient matrix given by

$$\mathbf{K}_\epsilon := \left[\begin{array}{c|c} K_P(t_j - \epsilon, t_i) & \partial_2 K_P(t_j - \epsilon, t_i) \\ \hline \frac{1}{2\epsilon}\big(K_P(t_j + \epsilon, t_i) - K_P(t_j - \epsilon, t_i)\big) & \frac{1}{2\epsilon}\big(\partial_2 K_P(t_j + \epsilon, t_i) - \partial_2 K_P(t_j - \epsilon, t_i)\big) \end{array}\right].$$

That is, the equations (2.2.3) can be written as

$$\mathbf{K}_\epsilon \begin{bmatrix} | \\ \alpha_\epsilon \\ | \\ | \\ \beta_\epsilon \\ | \end{bmatrix} = \begin{bmatrix} h(t_1-\epsilon) \\ \vdots \\ h(t_M-\epsilon) \\ \frac{1}{2\epsilon}(h(t_1+\epsilon)-h(t_1-\epsilon)) \\ \vdots \\ \frac{1}{2\epsilon}(h(t_M+\epsilon)-h(t_M-\epsilon)) \end{bmatrix}.$$

We first show that the matrix $\mathbf{K}_\epsilon$ is invertible for all ϵ sufficiently small. Note that as $\epsilon \to 0$ the matrix $\mathbf{K}_\epsilon$ converges to

$$\mathbf{K} := \left[\begin{array}{c|c} K_P(t_j,t_i) & \partial_2 K_P(t_j,t_i) \\ \hline \partial_1 K_P(t_j,t_i) & \partial_1\partial_2 K_P(t_j,t_i) \end{array}\right] = \int v(s)v(s)^T dP(s),$$

which is positive definite by Independence. Since the entries of $\mathbf{K}_\epsilon$ converge to the entries of $\mathbf{K}$, there is a $\Delta > 0$ such that $\mathbf{K}_\epsilon$ is invertible for all $\epsilon \in (0,\Delta)$. Moreover, $\mathbf{K}_\epsilon^{-1}$ converges to $\mathbf{K}^{-1}$ as $\epsilon \to 0$ and for all $\epsilon < \Delta$, the coefficients are uniquely defined as

$$\begin{bmatrix} | \\ \alpha_\epsilon \\ | \\ | \\ \beta_\epsilon \\ | \end{bmatrix} = \mathbf{K}_\epsilon^{-1} \begin{bmatrix} h(t_1-\epsilon) \\ \vdots \\ h(t_M-\epsilon) \\ \frac{1}{2\epsilon}(h(t_1+\epsilon)-h(t_1-\epsilon)) \\ \vdots \\ \frac{1}{2\epsilon}(h(t_M+\epsilon)-h(t_M-\epsilon)) \end{bmatrix}. \tag{2.2.4}$$

We denote the corresponding function by

$$\tilde{Q}_\epsilon(t) := \sum_{i=1}^{M} \alpha_\epsilon^{[i]} K_P(t,t_i) + \beta_\epsilon^{[i]} \partial_2 K_P(t,t_i).$$

Before we construct $\tilde{Q}(t)$, we take a moment to establish the following remarkable consequences of the Determinantal condition. For all $\epsilon > 0$ sufficiently small the following hold:

(a). $\tilde{Q}_\epsilon(t) = h(t)$ only at the points $t_1-\epsilon, t_1+\epsilon, \ldots, t_M-\epsilon, t_M+\epsilon$.

(b). These points $t_1-\epsilon, t_1+\epsilon, \ldots, t_M-\epsilon, t_M+\epsilon$ are nodal zeros of $\tilde{Q}_\epsilon(t) - h(t)$.

We adapted the proofs of (a) and (b) (with nontrivial modification) from the proofs of Theorem 1.6.1 and Theorem 1.6.2 of [66].

Proof of (a). Suppose for the sake of contradiction that there is a $\tau \in [-T, T]$ such that $\tilde{Q}_\epsilon(\tau) = h(\tau)$ and $\tau \notin \{t_1 - \epsilon, t_1 + \epsilon, \dots, t_M - \epsilon, t_M + \epsilon\}$. Then we have the system of $2M$ linear equations

$$\begin{aligned}\frac{\tilde{Q}_\epsilon(t_j - \epsilon)}{h(t_j - \epsilon)} - \frac{\tilde{Q}_\epsilon(\tau)}{h(\tau)} &= 0 \quad j = 1, \dots, M\\ \frac{\tilde{Q}_\epsilon(t_j + \epsilon)}{h(t_j + \epsilon)} - \frac{\tilde{Q}_\epsilon(\tau)}{h(\tau)} &= 0 \quad j = 1, \dots, M.\end{aligned}$$

Rewriting this in matrix form, the coefficient vector $\begin{bmatrix}\alpha_\epsilon & \beta_\epsilon\end{bmatrix} = \begin{bmatrix}\alpha_\epsilon^{[1]} & \cdots & \alpha_\epsilon^{[M]} & \beta_\epsilon^{[1]} & \cdots & \beta_\epsilon^{[M]}\end{bmatrix}$ of $\tilde{Q}_\epsilon$ satisfies

$$\begin{bmatrix}\alpha_\epsilon & \beta_\epsilon\end{bmatrix}\Big(\kappa(t_1 - \epsilon) - \kappa(\tau) \quad \kappa(t_1 + \epsilon) - \kappa(\tau) \quad \dots \quad \kappa(t_M + \epsilon) - \kappa(\tau)\Big) = \begin{bmatrix}0 & \dots & 0\end{bmatrix}. \tag{2.2.5}$$

By Lemma 2.2.3 applied to the $2M + 1$ vectors $v_1 = \kappa(t_1 - \epsilon), \dots, v_{2M} = \kappa(t_M + \epsilon)$, and $v_0 = \kappa(\tau)$, the matrix for the system of equations (2.2.5) is nonsingular if and only if the following matrix is nonsingular:

$$\begin{bmatrix}\kappa(t_1 - \epsilon) & \dots & \kappa(t_M + \epsilon) & \kappa(\tau)\\ 1 & \dots & 1 & 1\end{bmatrix} = \Lambda(t_1 - \epsilon, \dots, t_M + \epsilon, \tau).$$

However, this is nonsingular by the DETERMINANTAL condition. This gives us the contradiction that completes the proof of part (a).

Proof of (b). Suppose for the sake of contradiction that $\tilde{Q}_\epsilon(t) - h(t)$ has $N_1 < 2M$ nodal zeros and $N_0 = 2M - N_1$ non-nodal zeros. Denote the nodal zeros by $\{\tau_1, ..., \tau_{N_1}\}$, and denote the non-nodal zeros by $z_1, \dots, z_{N_0}$. In what follows, we obtain a contradiction by doubling the non-nodal zeros of $\tilde{Q}_\epsilon(t) - h(t)$. We do this by constructing a certain generalized polynomial $u(t)$ and adding a small multiple of it to $\tilde{Q}_\epsilon(t) - h(t)$.

We divide the non-nodal zeros into groups according to whether $\tilde{Q}_\epsilon(t) - h(t)$ is positive or negative in a small neighborhood around the zero; define

$$\mathcal{I}^- := \{i \mid \tilde{Q}_\epsilon \leq w \text{ near } z_i\} \quad \text{and} \quad \mathcal{I}^+ := \{i \mid \tilde{Q}_\epsilon \geq w \text{ near } z_i\}.$$

We first show that there are coefficients $a_0, \dots, a_M$, and $b_1, \dots, b_M$ such that the polynomial

$$u(t) = \sum_{i=1}^{M} a_i K_P(t, t_i) + \sum_{i=1}^{M} b_i \partial_2 K_P(t, t_i) + a_0 h(t)$$

satisfies the system of equations

$$\begin{aligned}u(z_j) &= +1 \quad j \in \mathcal{I}^-\\ u(z_j) &= -1 \quad j \in \mathcal{I}^+\\ u(\tau_i) &= 0 \quad i = 1, \dots, N_1\\ u(\tau) &= 0,\end{aligned} \tag{2.2.6}$$

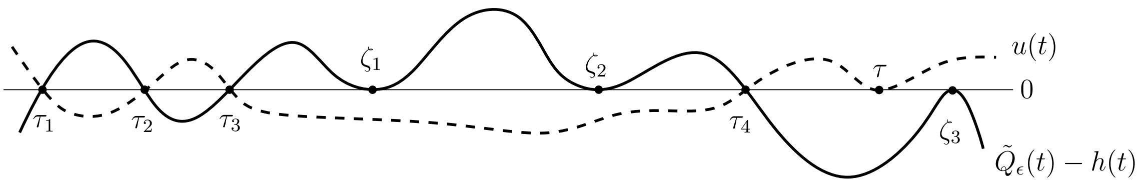

Figure 2.4: The points $\{\tau_1, \tau_2, \tau_3, \tau_4\}$ are nodal zeros of $\tilde{Q}_\epsilon(t) - h(t)$, and the points $\{\zeta_1, \zeta_2, \zeta_3\}$ are non-nodal zeros. The function $u(t)$ has the appropriate sign so that $\tilde{Q}_\epsilon(t) - h(t) + \delta u(t)$ retains nodal zeros at τ_i, and obtains two zeros in the vicinity of each ζ_i.

where τ is some arbitrary additional point. The matrix for this system is

$$\mathbf{W}\begin{pmatrix} \kappa(z_1)^T & 1 \\ \vdots & \\ \kappa(z_{N_0})^T & 1 \\ \kappa(\tau_1)^T & 1 \\ \vdots & \\ \kappa(\tau_{N_1})^T & 1 \\ \kappa(\tau) & 1 \end{pmatrix}$$

where $\mathbf{W} = \operatorname{diag}\big(h(z_1), \ldots, h(z_{N_0}), h(\tau_1), \ldots, h(\tau_{N_1}), h(\tau)\big)$. This matrix is invertible by Determinantal since the nodal and non-nodal zeros of $\tilde{Q}_\epsilon(t) - h(t)$ are given by $t_1 - \epsilon, \ldots, t_M + \epsilon$. Hence there is a solution to the system (2.2.6).

Now consider the function

$$U^\delta(t) = \tilde{Q}_\epsilon(t) + \delta u(t) = \sum_{i=1}^{M}[\alpha_\epsilon^{[i]} + \delta a_i]K_P(t, t_i) + \sum_{i=1}^{M}[\beta_\epsilon^{[i]} + \delta b_i]\partial_2 K_P(t, t_i) + \delta a_0 h(t)$$

where $\delta > 0$. By construction, $u(\tau_i) = 0$, so $U^\delta(t) - h(t)$ has nodal zeros at $\tau_1, \ldots, \tau_{N_1}$. We can choose δ small enough so that $U^\delta(t) - h(t)$ vanishes twice in the vicinity of each z_i. This means that $U^\delta(t) - h(t)$ has $2M + N_0$ zeros. Assuming $N_0 > 0$, select a subset of these zeros $p_1 < \ldots < p_{2M+1}$ such that there are two in each interval $[t_i - \rho, t_i + \rho]$. This is possible if $\epsilon < \rho$ and δ is sufficiently small. We have the system of $2M + 1$ equations

$$\begin{aligned} \sum_{i=1}^{M}[\alpha_\epsilon^{[i]} + \delta a_i]K_P(p_1, t_i) + \sum_{i=1}^{M}[\beta_\epsilon^{[i]} + \delta b_i]\partial_2 K_P(p_1, t_i) &= (1 - \delta a_0)h(\tau) \\ &\vdots \\ \sum_{i=1}^{M}[\alpha_\epsilon^{[i]} + \delta a_i]K_P(p_{2M+1}, t_i) + \sum_{i=1}^{M}[\beta_\epsilon^{[i]} + \delta b_i]\partial_2 K_P(p_{2M+1}, t_i) &= (1 - \delta a_0)h(\tau). \end{aligned}$$

Subtracting the last equation from each of the first $2M$ equations, we find that

$$(\alpha_\epsilon^{[1]} + \delta a_1, \ldots, \beta_\epsilon^{[M]} + \delta b_M)\begin{pmatrix} \kappa(p_1) - \kappa(p_{2M+1}) & \ldots & \kappa(p_{2M}) - \kappa(p_{2M+1}) \end{pmatrix} = (0, \ldots, 0).$$

This matrix is nonsingular by Lemma 2.2.3 combined with the Determinantal condition. This contradiction implies that $N_0 = 0$. This completes the proof of (b).

We now complete the proof by constructing $\tilde{Q}(t)$ from $\tilde{Q}_\epsilon(t)$ by sending $\epsilon \to 0$. Note that the coefficients $\alpha_\epsilon, \beta_\epsilon$ converge as $\epsilon \to 0$ since the right hand side of equation (2.2.4) converges to

$$\mathbf{K}^{-1}\begin{bmatrix} h(t_1) \\ \vdots \\ h(t_M) \\ h'(t_1) \\ \vdots \\ h'(t_M) \end{bmatrix} = \begin{bmatrix} | \\ \alpha \\ | \\ | \\ \beta \\ | \end{bmatrix}.$$

We denote the limiting function by

$$\tilde{Q}(t) = \sum_{i=1}^{M} \alpha_i K_P(t, t_i) + \sum_{i=1}^{M} \beta_i \partial_2 K_P(t, t_i). \tag{2.2.7}$$

We conclude that $h(t) - \tilde{Q}(t)$ does not change sign at the t_i since $h(t) - \tilde{Q}_\epsilon(t)$ changes sign only at $t_i \pm \epsilon$.

We now show that the limiting process does not introduce any additional zeros of $h(t) - \tilde{Q}(t)$. Suppose $\tilde{Q}(t)$ does touch $h(t)$ at some $\tau_1 \in [-T, T]$ with $\tau_1 \neq t_i$ for any $i = 1, ..., M$. Since $h(t) - \tilde{Q}(t)$ does not change sign, the points $t_1, \ldots, t_M, \tau_1$ are non-nodal zeros of $h(t) - \tilde{Q}(t)$. We find a contradiction by constructing a polynomial with two nodal zeros in the vicinity of each of these $M + 1$ points (but possibly only one nodal zero in the vicinity of τ_1 if $\tau_1 = T$ or $\tau_1 = -T$).

For sufficiently small $\gamma > 0$, the polynomial

$$W_\gamma(t) = \tilde{Q}(t) + \gamma h(t)$$

attains the value $h(t)$ twice in the vicinity of each t_i and twice in the vicinity of τ_1. In other words there exist $p_1 < \ldots < p_{2M+2}$ such that $W_\gamma(p_i) = h(p_i)$. Therefore

$$\tilde{Q}(p_i) = (1 - \gamma)h(p_i) \quad \text{for} \quad i = 1, \ldots, 2M + 2,$$

and so $\frac{\tilde{Q}(p_i)}{h(p_i)} - \frac{\tilde{Q}(p_{2M+1})}{h(p_{2M+1})} = 0$ for $i = 1, 2, ..., 2M$. Thus, the coefficient vector for the polynomial $\tilde{Q}(t)$ lies in the left nullspace of the matrix

$$\begin{pmatrix} \kappa(p_1) - \kappa(p_{2M+1}) & \ldots & \kappa(p_{2M}) - \kappa(p_{2M+1}) \end{pmatrix}.$$

However, this matrix is nonsingular by Lemma 2.2.3 and the Determinantal condition.

Collecting our results, we have proven that $\tilde{Q}(t) - h(t) = 0$ if and only if $t = t_i$ and that $\tilde{Q}(t) - h(t)$ does not change sign when t passes through t_i. Therefore one of the following is true

$$h(t) \geq \tilde{Q}(t) \quad \text{or} \quad \tilde{Q}(t) \geq h(t)$$

with equality iff $t = t_i$. In the first case, $Q(t) = \tilde{Q}(t)$ fulfills the prescriptions (2.2.1) with

$$q(t) = \sum_{i=1}^{M} \alpha_i \psi(s, t_i) + \beta_i \frac{d}{dt_i} \psi(s, t_i).$$

In the second case, $Q(t) = 2h(t) - \tilde{Q}(t)$ satisfies (2.2.1) with

$$q(t) = 2 - \sum_{i=1}^{M} \alpha_i \psi(s, t_i) + \beta_i \frac{d}{dt_i} \psi(s, t_i).$$

Proof of Theorem 2.1.1

Integrability and Positivity naturally hold for the Gaussian point spread function $\psi(s,t) = e^{-(s-t)^2}$. Independence holds because $\psi(s,t_1), \ldots, \psi(s,t_M)$ together with their derivatives $\partial_2\psi(s,t_1), \ldots, \partial_2\psi(s,t_M)$ form a T-system (see for example [66]). This means that for any $s_1 < \ldots < s_{2M} \in \mathbb{R}$,

$$\left| v(s_1) \ldots v(s_{2M}) \right| \neq 0,$$

and the determinant always takes the same sign. Therefore, by an integral version of the Cauchy-Binet formula for the determinant (cf. [65]),

$$\left| \int v(s) v(s)^T dP(s) \right| = (2M)! \int_{s_1 < \ldots < s_{2M}} \left| v(s_1) \ldots v(s_{2M}) \right| \begin{vmatrix} v(s_1)^T \\ \vdots \\ v(s_{2M})^T \end{vmatrix} dP(s_1) \ldots dP(s_{2M}) \neq 0.$$

To establish the Determinantal condition, we prove the slightly stronger statement:

$$|\Lambda(p_1, \ldots, p_{2M+1})| = \left| \int \begin{bmatrix} v(s) \\ 1 \end{bmatrix} \begin{bmatrix} \frac{\psi(s,p_1)}{h(p_1)} & \ldots & \frac{\psi(s,p_{2M+1})}{h(p_{2M+1})} \end{bmatrix} dP(s) \right| \neq 0 \tag{2.2.8}$$

for any distinct $p_1, \ldots, p_{2M+1}$. When $p_1, \ldots, p_{2M+1}$ are restricted so that two points p_i, p_j lie in each ball $(t_k - \rho, t_k + \rho)$, we recover the statement of Determinantal.

We prove (2.2.8) with the following key lemma.

Lemma 2.2.4. *For any $s_1 < \ldots < s_{2M+1}$ and $t_1 < \ldots < t_M$,*

$$\begin{vmatrix} e^{-(s_1-t_1)^2} & \ldots & e^{-(s_{2M+1}-t_1)^2} \\ -(s_1-t_1)e^{-(s_1-t_1)^2} & \ldots & -(s_{2M+1}-t_1)e^{-(s_{2M+1}-t_1)^2} \\ \vdots & & \vdots \\ e^{-(s_1-t_M)^2} & \ldots & e^{-(s_{2M+1}-t_M)^2} \\ -(s_1-t_M)e^{-(s_1-t_M)^2} & \ldots & -(s_{2M+1}-t_M)e^{-(s_{2M+1}-t_M)^2} \\ 1 & \ldots & 1 \end{vmatrix} \neq 0.$$

Before proving this lemma, we show how it can be used to prove (2.2.8). By Lemma 2.2.4, we know in particular that for any $s_1 < \cdots < s_{2M+1}$,

$$\det \begin{bmatrix} v(s_1) & \cdots & v(s_{2M+1}) \\ 1 & \cdots & 1 \end{bmatrix} \neq 0$$

and is always the same sign. Moreover, for any $s_1 < \cdots < s_{2M+1}$, and any $p_1 < \ldots < p_{2M+1}$,

$$\det \begin{bmatrix} \psi(s_1, p_1) & \ldots & \psi(s_1, p_{2M+1}) \\ \vdots & & \\ \psi(s_{2M+1}, p_1) & \ldots & \psi(s_{2M+1}, p_{2M+1}) \end{bmatrix} > 0.$$

Any function with this property is called *totally positive* and it is well known that the Gaussian kernel is totally positive [66]. Now, to show that DETERMINANTAL holds for the finite sampling measure P, we use an integral version of the Cauchy-Binet formula for the determinant:

$$\left| \int \begin{bmatrix} v(s) \\ 1 \end{bmatrix} \begin{bmatrix} \frac{\psi(s,p_1)}{h(p_1)} & \ldots & \frac{\psi(s,p_{2M+1})}{h(p_{2M+1})} \end{bmatrix} dP(s) \right| =$$

$$= (2M+1)! \int\limits_{s_1 < \cdots < s_{2M+1}} \begin{vmatrix} v(s_1) & \cdots & v(s_{2M+1}) \\ 1 & \cdots & 1 \end{vmatrix} \begin{vmatrix} \frac{\psi(s_1,p_1)}{h(p_1)} & \ldots & \frac{\psi(s_1,p_{2M+1})}{h(p_{2M+1})} \\ \vdots & & \\ \frac{\psi(s_{2M+1},p_1)}{h(p_1)} & \ldots & \frac{\psi(s_{2M+1},p_{2M+1})}{h(p_{2M+1})} \end{vmatrix} dP(s_1) \ldots dP(s_{2M+1}).$$

The integral is nonzero since all integrands are nonzero and have the same sign. This proves (2.2.8).

Proof of Lemma 2.2.4. Multiplying the $2i-1$ and $2i$-th row by $e^{t_i^2}$ and the i-th column by $e^{s_i^2}$, and subtracting t_i times the $2i-1$-th row from the $2i$-th row, we obtain that we equivalently have to show that

$$\begin{vmatrix} e^{s_1 t_1} & e^{s_2 t_1} & \ldots & e^{s_{2M+1} t_1} \\ s_1 e^{s_1 t_1} & s_2 e^{s_2 t_1} & \ldots & s_{2M+1} e^{s_k t_1} \\ e^{s_1 t_2} & e^{s_2 t_2} & \ldots & e^{s_{2M+1} t_2} \\ \vdots & & & \\ e^{s_1 t_M} & e^{s_2 t_M} & \ldots & e^{s_{2M+1} t_M} \\ s_1 e^{s_1 t_M} & s_2 e^{s_2 t_M} & \ldots & s_{2M+1} e^{s_{2M+1} t_M} \\ e^{s_1^2} & e^{s_2^2} & \ldots & e^{s_{2M+1}^2} \end{vmatrix} \neq 0.$$

The above matrix has a vanishing determinant if and only if there exists a nonzero vector

$$(a_1, b_1, ..., a_M, b_M, a_{M+1})$$

in its left null space. This vector has to have nonzero last coordinate since by Example 1.1.5. in [66], the Gaussian kernel is extended totally positive and therefore the upper $2M \times 2M$ submatrix has a nonzero determinant. Therefore, we assume that $a_{M+1} = 1$. Thus, the matrix above has a vanishing determinant if and only if the function

$$\sum_{i=1}^{M} (a_i + b_i s) e^{t_i s} + e^{s^2} \tag{2.2.9}$$

has at least the $2M+1$ zeros $s_1 < s_2 < ... < s_{2M+1}$. Lemma 2.2.5, applied to $r = M$ and $d_1 = \cdots = d_M = 1$, establishes that this is impossible. To complete the proof of Lemma 2.2.4, it remains to state and prove Lemma 2.2.5. □

Remark. The inclusion of the derivatives is essential for the shifted Gaussians to form a T-system together with the constant function 1. In particular, following the same logic as in the proof of Lemma 2.2.4, we find that $\{1, e^{(s-t_1)^2}, \dots, e^{(s-t_M)^2}\}$ form a T-system if and only if the function

$$\sum_{i=1}^{M} a_i e^{t_i s} + e^{s^2}$$

has at most M zeros. However, for $M = 3$ the function has 4 zeros if we select $a_1 = -3$, $t_1 = 1$, $a_2 = 7$, $t_2 = 0$, $a_3 = -5$, $t_3 = -1$.

Lemma 2.2.5. *Let $d_1, ..., d_r \in \mathbb{N}$. The function*

$$\phi_{d_1,\dots,d_r}(s) = \sum_{i=1}^{r} (a_{i0} + a_{i1}s + \cdots + a_{i(2d_i-1)} s^{2d_i-1}) e^{t_i s} + e^{s^2}$$

has at most $2(d_1 + \cdots + d_r)$ zeros.

Proof. We are going to show that $\phi_{d_1,\dots,d_r}(s)$ has at most $2(d_1 + \cdots + d_r)$ zeros as follows. Let

$$g_0(s) = \phi_{d_1,\dots,d_r}(s).$$

For $k = 1, ..., d_1 + \cdots + d_r$, let

$$g_k(s) = \begin{cases} \frac{d^2}{ds^2}\left[g_{k-1}(s) e^{(-t_j + t_1 + \cdots + t_{j-1})s}\right], & \text{if } k = d_1 + \cdots + d_{j-1} + 1 \text{ for some } j, \\ \frac{d^2}{ds^2}\left[g_{k-1}(s)\right], & \text{otherwise.} \end{cases} \tag{2.2.10}$$

If we show that $g_{d_1+\dots+d_r}(s)$ has no zeros, then, $g_{d_1+\dots+d_r-1}(s)$ has at most two zeros, counting with multiplicity. By induction, it will follow that $g_0(s)$ has at most $2(d_1 + \cdots + d_r)$ zeros, counting with multiplicity. Note that if $d_1 + \cdots + d_{j-1} \le k < d_1 + \cdots + d_{j-1} + d_j$, then

$$\begin{aligned} g_k(s) =& (\tilde{a}_{j,2(k-d_1+\dots+d_{j-1})} + \cdots + \tilde{a}_{j,(2d_j-1)} s^{2d_j-1-2(k-d_1+\dots+d_{j-1})}) + \\ &+ \sum_{i=j+1}^{r} (\tilde{a}_{i0} + \cdots + \tilde{a}_{i(2d_i-1)} s^{2d_i-1}) e^{(t_i-(t_1+\dots+t_{j-1}))r} + c f_i(r) e^{r^2} \end{aligned}$$

where $c > 0$ is a constant and $r := s - c_i$. We are going to show that $f_i(r)$ is a sum of squares polynomial such that one of the squares is a positive constant. This would mean that $g_k(s) = f_k(s) e^{s^2}$ has no zeros.

Denote

$$\begin{aligned} p_0(s) &= 1 \\ p_1(s) &= 2s \\ &\vdots \\ p_i(s) &= 2s p_{i-1}(s - c_i) + p'_{i-1}(s - c_i), \end{aligned}$$

where $c_1, ..., c_k$ are constants. It follows by induction that the degree of $p_i(s)$ is $\deg(p_i) = i$ and the leading coefficient of $p_i(s)$ is 2^i.

We will show by induction that

$$\begin{aligned} f_i(s) &= p_i(s)^2 + \frac{1}{2}p_i'(s)^2 + \cdots + \frac{1}{2^i i!}p_i^{(i)}(s)^2 \\ &= \sum_{j=0}^{i} \frac{1}{2^j j!} p_i^{(j)}(s)^2. \end{aligned}$$

When $i = 0$, we have that $f_0(s) = 1$ and $\sum_{j=0}^{0} \frac{1}{2^j j!} p_0^{(j)}(s)^2 = 1$. We are going to prove the general statement by induction. Suppose the statement is true for $i - 1$. By the relationship (2.2.10), we have

$$\begin{aligned} f_i(s)e^{s^2} =& \frac{d^2}{ds^2}\left[e^{s^2} f_{i-1}(s - c_i)\right] = \frac{d^2}{ds^2}\left[e^{s^2} \sum_{j=0}^{i-1} \frac{1}{2^j j!} p_{i-1}^{(j)}(s - c_i)^2\right] \qquad (2.2.11) \\ =& \sum_{j=0}^{i-1} \frac{e^{s^2}}{2^j j!}\Big\{2p_{i-1}^{(j+2)}(s - c_i)p_{i-1}^{(j)}(s - c_i) + 2p_{i-1}^{(j+1)}(s - c_i)^2 \\ &+ (4s^2 + 2)p_{i-1}^{(j)}(s - c_i)^2 + 8sp_{i-1}^{(j)}(s - c_i)p_{i-1}^{(j+1)}(s - c_i)\Big\} \end{aligned}$$

We need to show that this expression is equal to $e^{s^2}(\sum_{j=0}^{i} \frac{p_i^{(j)}(s)^2}{2^j j!})$. Since

$$p_i(s) = 2sp_{i-1}(s - c_i) + p_{i-1}'(s - c_i),$$

it follows by induction that $p_i^{(j)}(s) = 2jp_{i-1}^{(j-1)}(s - c_i) + 2sp_{i-1}^{(j)}(s - c_i) + p_{i-1}^{(j+1)}(s - c_i)$. Therefore we obtain

$$\begin{aligned} e^{s^2}(\sum_{j=0}^{i} \frac{p_i^{(j)}(s)^2}{2^j j!}) =& e^{s^2} \sum_{j=0}^{i} \frac{1}{2^j j!}\left[2jp_{i-1}^{(j-1)}(s - c_i) + 2sp_{i-1}^{(j)}(s - c_i) + p_{i-1}^{(j+1)}(s - c_i)\right]^2. \\ =& e^{s^2} \sum_{j=0}^{i} \frac{1}{2^j j!}\Big[4j^2p_{i-1}^{(j-1)}(s - c_i)^2 + 4s^2p_{i-1}^{(j)}(s - c_I)^2 + p_{i-1}^{(j+1)}(s - c_i)^2 + \\ &+ 8jsp_{i-1}^{(j-1)}(s - c_i)p_{i-1}^{(j)}(s - c_i) + \\ &+ 4sp_{i-1}^{(j)}(s - c_i)p_{i-1}^{(j+1)} + 4jp_{i-1}^{(j-1)}(s - c_i)p_{i-1}^{(j+1)}(s - c_i)\Big] \end{aligned} \qquad (2.2.12)$$

There are four types of terms in the sums (2.2.11) and (2.2.12):

$$p_{i-1}^{(j)}(s - c_i)^2, \quad s^2p_{i-1}^{(j)}(s - c_i)^2, \quad p_{i-1}^{(j-1)}(s - c_i)p_{i-1}^{(j)}(s - c_i), \quad \text{and} \quad sp_{i-1}^{(j-1)}(s - c_i)p_{i-1}^{(j)}(s - c_i).$$

For a fixed $j \in \{0, 1, ..., i + 1\}$, it is easy to check that the coefficients in front of each of these terms in (2.2.11) and (2.2.12) are equal. Therefore,

$$f_i(s) = p_i(s)^2 + \frac{1}{2}p_i'(s)^2 + \cdots + \frac{1}{2^i i!}p_i^{(i)}(s)^2$$

$$= \sum_{j=0}^{i} \frac{1}{2^j j!} p_i^{(j)}(s)^2$$

Note that since $\deg(p_i) = i$, then, $p_i^{(i)}(s)$ equals the leading coefficient of $p_i(s)$, which, as we discussed above, equals 2^i. Therefore, the term $\frac{1}{2^i i!} p_i^{(i)}(s)^2 = 2^i i!$. Thus, one of the squares in $f_i(s)$ is a positive number, so $f_i(s) > 0$ for all s. □

2.3 Numerical experiments

In this section we present the results of several numerical experiments to complement our theoretical results. To allow for potentially noisy observations, we solve the constrained least squares problem

$$\begin{aligned} \underset{\mu \geq 0}{\text{minimize}} \quad & \sum_{i=1}^{n} \left(\int \psi(s_i, t) d\mu(t) - x(s_i) \right)^2 \\ \text{subject to} \quad & \int h(t)\mu(dt) \leq \tau \end{aligned} \tag{2.3.1}$$

using the conditional gradient method proposed in [23].

Reweighing matters for source localization

Our first numerical experiment provides evidence that weighting by $h(t)$ helps recover point sources near the border of the image. This matches our intuition: near the border, the mass of an observed point-source is smaller than if it were measured in the center of the image. Hence, if we didn't weight the candidate locations, sources that are close to the edge of the image would be beneficial to add to the representation.

We simulate two populations of images, one with point sources located away from the image boundary, and one with point sources located near the image boundary. For each population of images, we solve (2.3.1) with $h(t) = \int \psi(s,t) dP(s)$ (weighted) and with $h(t) = 1$ (unweighted). We find that the solutions to (2.3.1) recover the true point sources more accurately with $h(t) = \int \psi(s,t) dP(s)$.

We use the same procedure for computing accuracy as in [98]. Namely we match true point sources to estimated point courses and compute the *F-score* of the match. To describe this procedure in detail, we compute the F-score by solving a bipartite graph matching problem. In particular, we form the bipartite graph with an edge between t_i and $\hat{t}_j$ for all i, j such that $\|t_i - \hat{t}_j\| < r$, where $r > 0$ is a tolerance parameter, and $\hat{t}_1, \ldots, \hat{t}_N$ are the estimated point sources. Then we greedily select edges from this graph under the constraint that no two selected edges can share the same vertex; that is, no t_i can be paired with two $\hat{t}_j, \hat{t}_k$ or vice versa. Finally, the $\hat{t}_i$ successfully paired with some t_j are categorized as true positives, and we denote their number by T_P. The number of false negatives is $F_N = M - T_P$, and the number of false positives is $N - T_P$. The precision and recall are then $P = \frac{T_P}{T_P + F_N}$, and $R = \frac{T_P}{T_P + F_P}$ respectively, and the F-score is the harmonic mean:

$$F = \frac{2PR}{P+R}.$$

We find a match by greedily pairing points of $\{\tau_1, \ldots, \tau_N\}$ to elements of $\{t_1, \ldots, t_M\}$, and a tolerance radius $r > 0$ upper bounds the allow distance between any potential pairs. To emphasize the dependence on r, we sometimes write $F(r)$ for the F-score.

Both populations contain 100 images simulated using the Gaussian point spread function

$$\psi(s,t) = e^{-\frac{(s-t)^2}{\sigma^2}}$$

with $\sigma = 0.1$, and in both cases, the measurement set $\mathcal{S}$ is a dense uniform grid of $n = 100$ points covering $[0,1]$. The populations differ in how the point sources for each image are chosen. Each image in the first population has five points drawn uniformly in the interval $(.1, .9)$, while each image in the second population has a total of four point sources with two point sources in each of the two boundary regions $(0, .1)$ and $(.9, 1)$. In both cases we assign intensity of 1 to all point sources, and solve (2.3.1) using an optimal value of τ (chosen with a preliminary simulation).

The results are displayed in Figure 2.5. The left subplot shows that the F-scores are essentially the same for the weighted and unweighted problems when the point sources are away from the boundary. This is not surprising because when t is away from the border of the image, then $\int \psi(s,t)dP(s)$ is essentially a constant, independent of t. But when the point sources are near the boundary, the weighting matters and the F-scores are dramatically better as shown in the right subplot.

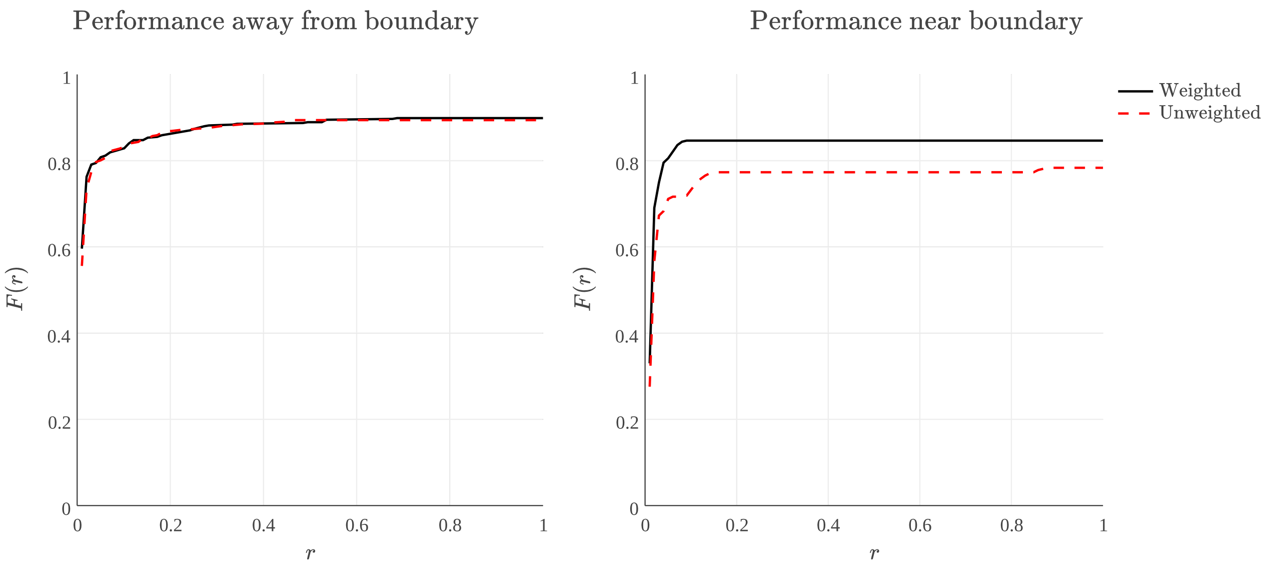

Figure 2.5: **Reweighing matters for source localization.** The two plots above compare the quality of solutions to the weighted problem (with $h(t) = \int \psi(s,t)dP(s)$) and the unweighted problem (with $h(t) = 1$). When point sources are away from the boundary (left plot), the performance is nearly identical. But when the point sources are near the boundary (right plot), the weighted method performs significantly better.

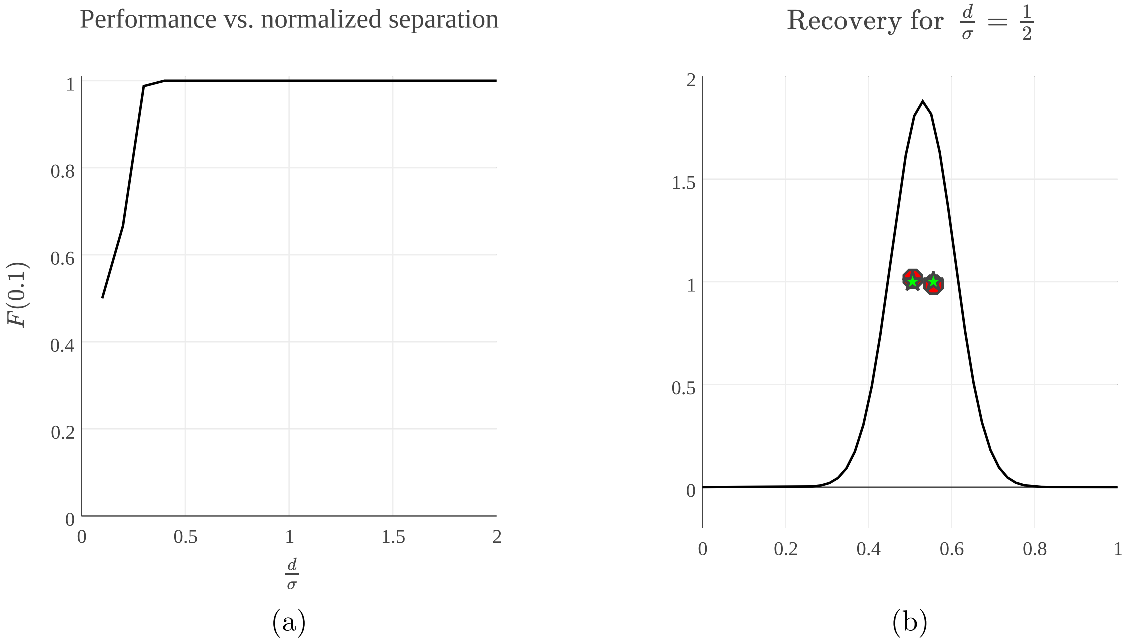

Figure 2.6: **Sensitivity to point-source separation.** (a) The F-score at tolerance radius $r = 0.1$ as a function of normalized separation $\frac{d}{\sigma}$. (b) The black trace shows an image for $\frac{d}{\sigma} = \frac{1}{2}$. The green stars show the locations (x-coordinate) and weights (y-coordinate) of the true point sources. The red dots show the recovered locations and weights.

Sensitivity to point-source separation

Our theoretical results assert that in the absence of noise the optimal solution of (2.1.3) recovers point sources with no minimum bound on the separation. In the following experiment, we explore the ability of (2.3.1) to recover pairs of points as a function of their separation. The setup is similar to the first numerical experiment. We use the Gaussian point spread function with $\sigma = 0.1$ as before, but here we observe only $n = 50$ samples. For each separation $d \in \{.1\sigma, .2\sigma, \ldots, 1.9\sigma, 2\sigma\}$, we simulate a population of 20 images containing two point sources separated by d. The point sources are chosen by picking a random point x away from the border of the image and placing two point sources at $x \pm \frac{d}{2}$. Again, each point source is assigned an intensity of 1, and we attempt to recover the locations of the point sources by solving (2.3.1).

In the left subplot of Figure 2.6 we plot F-score versus separation for the value of τ that produces the best F-scores. Note that we achieve near perfect recovery for separations greater than $\frac{\sigma}{4}$. The right subplot of Figure 2.6 shows the observations, true point sources, and estimated point sources for a separation of $\frac{d}{\sigma} = \frac{1}{2}$. Note the near perfect recovery in spite of the small separation.

Due to numerical issues, we cannot localize point sources with arbitrarily small $d > 0$. Indeed, the F-score for $\frac{d}{\sigma} < \frac{1}{4}$ is quite poor. This does not contradict our theory because numerical ill-conditioning is in effect adding noise to the recovery problem, and we expect that a separation condition will be necessary in the presence of noise.

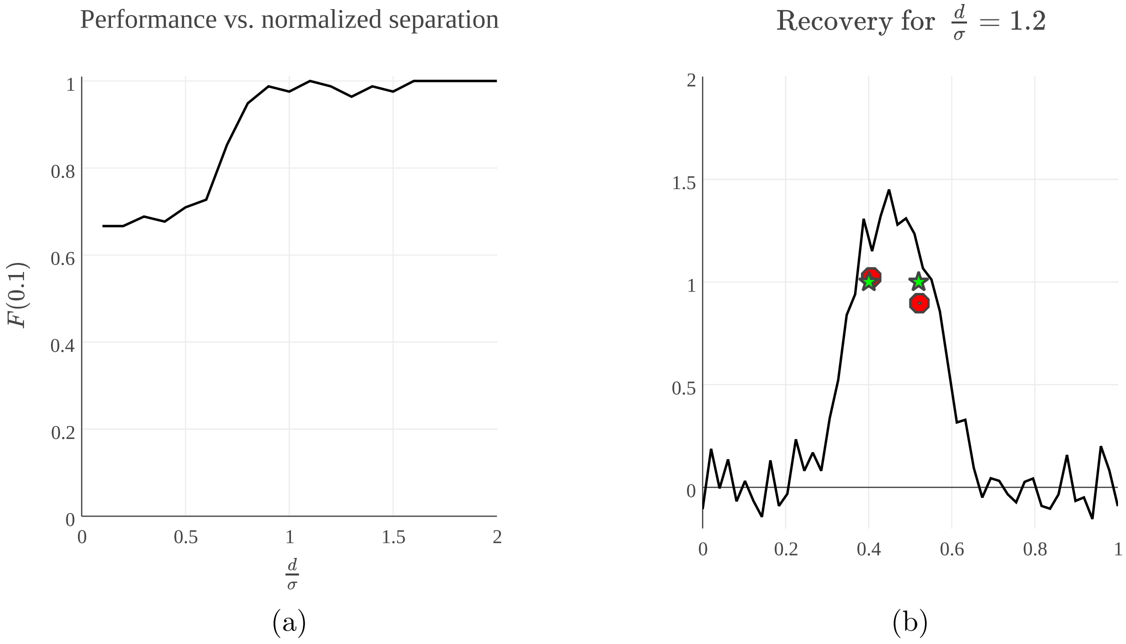

Figure 2.7: **Sensitivity to noise.** (a) The F-score at tolerance radius $r = 0.1$ as a function of normalized separation $\frac{d}{\sigma}$. (b) The black trace is the 50 pixel image we observe. The green stars show the locations (x-coordinate) and weights (y-coordinate) of the true point sources. The red dots show the recovered locations and weights.

Sensitivity to noise

Next, we investigate the performance of (2.3.1) in the presence of additive noise. The setup is identical to the previous numerical experiment, except that we add Gaussian noise to the observations. In particular, our noisy observations are

$$\{x(s_i) + \eta_i \mid s_i \in \mathcal{S}\}$$

where $\eta_i \sim \mathcal{N}(0, 0.1)$.

We measure the performance of (2.3.1) in Figure 2.7. Note that we achieve near-perfect recovery when $d > \sigma$. However, if $d < \sigma$ the F-scores are clearly worse than the noiseless case. Unsurprisingly, we observe that sources must be separated in order to recover their locations to reasonable precision. We defer an investigation of the dependence of the signal separation as a function of the signal-to-noise ratio to future work.

Extension to two-dimensions

Though our proof does not extend as is, we do expect generalizations of our recovery result to higher dimensional settings. The optimization problem (2.3.1) extends immediately to arbitrary dimensions, and we have observed that it performs quite well in practice. We demonstrate in Figure 2.8 the power of applying (2.3.1) to a high density fluorescence image in simulation. Figure 2.8 shows an image simulated with parameters specified by the Single Molecule Localization Microscopy challenge [20]. In this challenge, point sources are blurred

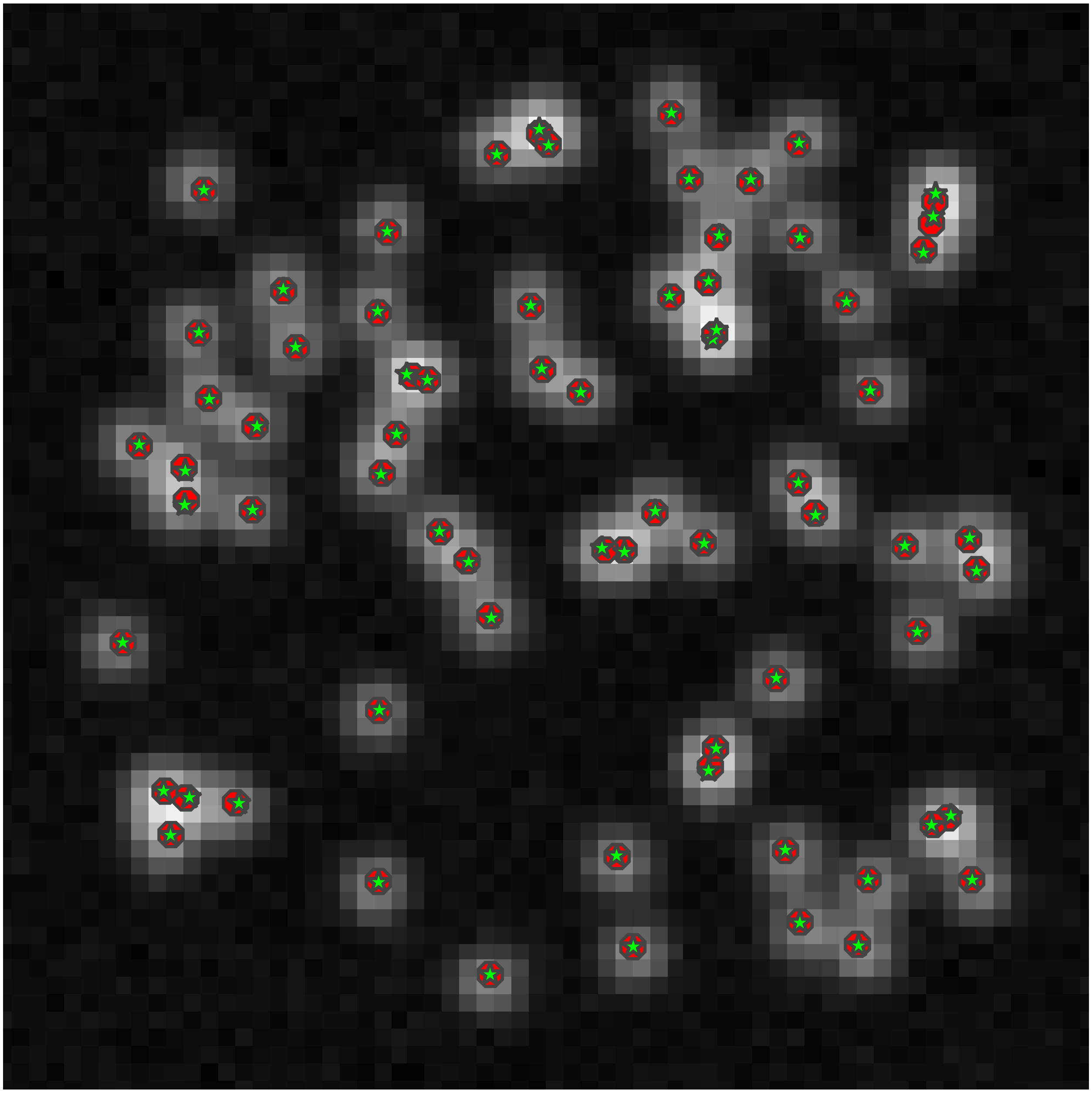

Figure 2.8: **High density single molecule imaging.** The green stars show the locations of a simulated collection point sources, and the greyscale background shows the noisy, pixelated point spread image. The red dots show the support of the measure-valued solution of (2.3.1).

by a Gaussian point-spread function and then corrupted by noise. The green stars show the true locations of a simulated collection of point sources, and the red dots show the support of the measure output by (2.3.1) applied to the greyscale image forming the background of Figure 2.8. The overlap between the true locations and estimated locations is near perfect with an F-score of 0.98 for a tolerance radius corresponding to one third of a pixel.

2.4 Conclusions and Future Work

In this chapter we have demonstrated that one can recover the centers of a nonnegative sum of Gaussians from a few samples by solving a convex optimization problem. This recovery is theoretically possible no matter how close the true centers are to one-another. We remark that similar results are true for recovering measures from their moments. Indeed, the atoms of a positive atomic measure can be recovered no matter how close together the atoms are, provided one observes twice the number of moments as there are atoms. Our work can be seen as a generalization of this result, applying generalized polynomials and the theory of Tchebycheff systems in place of properties of Vandermonde systems.

As we discussed in our numerical experiments, this work opens up several theoretical problems that would benefit from future investigation. We close with a very brief discussion of some of the possible extensions.

Noise Motivated by the fact that there is no separation condition in the absence of noise, it would be interesting to study how the required separation decays to zero as the noise level decreases. One of the key-advantages of using convex optimization for signal processing is that dual certificates generically give stability results, in the same way that Lagrange multipliers measure sensitivity in linear programming. Previous work on estimating line-spectra has shown that dual polynomials constructed for noiseless recovery extend to certify properties of estimation and localization in the presence of noise [26, 48, 111]. We believe that these methods should be directly applicable to our problem set-up.

Higher dimensions One logical extension is proving that the same results hold in higher dimensions. Most scientific and engineering applications of interest have point sources arising one to four dimensions, and we expect that some version of our results should hold in higher dimensions. Indeed, we believe a guarantee for recovery with no separation condition can be proven in higher dimensions with noiseless observations. However, it is not straightforward to extend our results to higher dimensions because the theory of Tchebycheff systems is only developed in one dimension. In particular, our approach using limits of polynomials does not directly generalize to higher dimensions.

Other point spread functions We have shown that our Conditions 2.1.2 hold for the Gaussian point spread function, which is commonly used in microscopy as an approximation to an Airy function. It will be very useful to show that they also hold for other point spread functions such as the Airy function and other common physical models. Our proof relied heavily on algebraic properties of the Gaussian, but there is a long, rich history of determinantal systems that may apply to generalize our result. In particular, works on properties of totally positive systems may be fruitful for such generalizations [3, 86].

Model mismatch in the point spread function Our analysis relies on perfect knowledge of the point spread function. In practice one never has an exact analytic expression for the point spread function. Aberrations in manufacturing and scattering media can lead to distortions in the image not properly captured by a forward model. It would be interesting to derive guarantees on recovery that assume only partial knowledge of the point spread function. Note that the optimization problem of searching both for the locations of the sources and for the associated wave-function is a blind deconvolution problem, and techniques from this well-studied problem could likely be extended to the super-resolution setting. If successful, such methods could have immediate practical impact when applied to denoising images in molecular, cellular, and astronomical imaging.

Chapter 3

The Alternating Descent Conditional Gradient Method

In this chapter we propose a variant of the classical conditional gradient method (CGM) for sparse inverse problems with differentiable observation models. Such models arise in many practical problems including superresolution, time-series modeling, and matrix completion. Our algorithm combines nonconvex and convex optimization techniques: we propose global conditional gradient steps alternating with nonconvex local search exploiting the differentiable observation model. This hybridization gives the theoretical global optimality guarantees and stopping conditions of convex optimization along with the performance and modeling flexibility associated with nonconvex optimization. Our experiments demonstrate that our technique achieves state-of-the-art results in several applications.

This chapter is joint work with Nicholas Boyd and Benjamin Recht. The content of this chapter has been submitted for publication under the title *The Alternating Descent Conditional Gradient Method for Sparse Inverse Problems* and is available on the arxiv http://arxiv.org/abs/1507.01562.

3.1 Introduction

A ubiquitous prior in modern statistical signal processing asserts that an observed signal is the noisy observation of a few weighted sources. In other words, compared to the entire dictionary of possible sources, the set of sources actually present is *sparse*. In the most abstract formulation of this prior, each source is chosen from a non-parametric dictionary, but in many cases of practical interest the sources are parameterized. Hence, solving the sparse inverse problem amounts to finding a collection of a few parameters and weights that adequately explains the observed signal.

As a concrete example, consider the idealized task of identifying the aircraft that lead to an observed radar signal. The sources are the aircraft themselves, and each is parameterized by, perhaps, its position and velocity relative to the radar detector. The sparse inverse problem is to recover the number of aircraft present, along with each of their parameters.

Any collection of weighted sources can be represented as a measure on the parameter space: each source corresponds to a single point mass at its corresponding parameter value.

We will call atomic measures supported on very few points *sparse* measures. When the parameter spaces are infinite—for example the set of all velocities and positions of aircraft—the space of sparse measures over such parameters is infinite-dimensional. This means that optimization problems searching for parsimonious explanations of the observed signal must operate over an infinite-dimensional space.

Many alternative formulations of the sparse inverse problem have been proposed to avoid the infinite-dimensional optimization required in the sparse measure setup. The most canonical and widely applicable approach is to form a discrete grid over the parameter space and restrict the search to measures supported on the grid. This restriction produces a finite-dimensional optimization problem [18, 78, 112]. In certain special cases, the infinite-dimensional optimization problem over measures can be reduced to a problem of moment estimation, and spectral techniques or semidefinite programming can be employed [53, 87, 113, 27]. More recently, in light of much of the work on compressed sensing and its generalizations, another proposal operates on atomic norms over data [35], opening other algorithmic possibilities.

While these finite-dimensional formulations are appealing, they all essentially treat the space of sources as an unstructured set, ignoring natural structure (such as differentiability) present in many applications. All three of these techniques have their individual drawbacks, as well. Gridding only works for very small parameter spaces, and introduces artifacts that often require heuristic post-processing [112]. Moment methods have limited applicability, are typically computationally expensive, and, moreover, are sensitive to noise and estimates of the number of sources. Finally, atomic norm techniques do not recover the parameters of the underlying signal, and as such are more naturally applied to denoising problems.

In this chapter, we argue that all of these issues can be alleviated by returning to the original formulation of the estimation problem as an optimization problem over the space of measures. Working with measures explicitly exposes the underlying parameter space, which allows us to consider algorithms that make local moves within parameter space. We demonstrate that operating on the infinite-dimensional space of measures is not only feasible algorithmically, but that the resulting algorithms outperform techniques based on gridding or moments on a variety of real-world signal processing tasks. We formalize a general approach to solving parametric sparse inverse problems via the conditional gradient method (CGM), also know as the Frank-Wolfe algorithm. In Section 3.3, we show how to augment the classical CGM with nonconvex local search exploiting structure in the parameter space. This hybrid scheme, which we call the alternating descent conditional gradient method (ADCG), enjoys both the rapid local convergence of nonconvex programming algorithms and the stability and global convergence guarantees associated with convex optimization. The theoretical guarantees are detailed in Section 3.5, where we bound the convergence rate of our algorithm and also guarantee that it can be run with bounded memory. Moreover, in Section 3.6 we demonstrate that our approach achieves state-of-the-art performance on a diverse set of examples.

Mathematical setup

In this subsection we formalize the sparse inverse problem as an optimization problem over measures and discuss a convex heuristic.

We assume the existence of an underlying collection of objects, called sources. Each source has a scalar weight w, and a parameter $t \in \Theta$. We require the parameter space be measurable (that is, come equiped with a σ-algebra) and amenable to local, derivative-based optimization. Some examples to keep in mind would be $\Theta = \mathbb{R}^p$ for some small p, or the sphere $\mathcal{S}^p$ considered as a differentiable manifold. An element t of the parameter space Θ may describe, for instance, the position, orientation, and polarization of a source. The weight w may encode the intensity of a source, or the distance of a source from the observation device. Our goal is to recover the number of sources present, along with their individual weights and parameters. We do not observe the sources directly, but instead are given a single, noisy observation in $\mathbb{R}^d$.

The observation model we use is completely specified by a function $\phi : \Theta \to \mathbb{R}^d$, which gives the d-dimensional observation of a single, unit-weight source parameterized by a point in Θ. A single source with parameter t and weight w generates the observation $w\phi(t) \in \mathbb{R}^d$: that is, the observation of a lone source is homogeneous of degree one in its weight. The observation of a lone source is homogeneous of degree one in its weight; that is, a single source with parameter t and weight w generates the observation $w\phi(t) \in \mathbb{R}^d$ Finally, we assume that the observation generated by a weighted collection of sources is additive. In other words, the (noise-free) observation of a weighted collection of sources, $\{(w_i, t_i)\}_{i=1}^M$, is simply

$$\sum_{i=1}^{M} w_i \phi(t_i) \in \mathbb{R}^d. \tag{3.1.1}$$

We refer to the collection $\{(w_i, t_i)\}_{i=1}^M$ as the *signal parameters*, and the vector $\sum_{i=1}^M w_i\phi(t_i) \in \mathbb{R}^d$ as the noise-free *observation*. We require ϕ to be bounded: $\|\phi(t)\|_2^2 \leq 1$ for all t, and further that ϕ be *differentiable* in t. Finally, let us emphasize that we make no further assumptions about ϕ: in particular it does *not* need to be linear or convex.

Our goal is to recover the true weighted collection of sources, $\{(\tilde{w}_i, \tilde{t}_i)\}_{i=1}^{\tilde{M}}$, from a single noisy observation:

$$y = \sum_{i=1}^{\tilde{M}} \tilde{w}_i \phi(\tilde{t}_i) + \nu.$$

Here ν is an additive noise term.

One approach would be to attempt to minimize a convex loss, ℓ, of the residual between the observed vector y and the expected output for an estimated collection of sources:

$$\underset{w,t,K}{\text{minimize}} \quad \ell\left(\sum_{i=1}^{M} w_i\phi(t_i) - y\right). \tag{3.1.2}$$

For example, when ℓ is the negative log-likelihood of the noise term ν, problem (3.1.2) corresponds to maximum-likelihood estimation of the true sources. Unfortunately, (3.1.2) is nonconvex in the variables w, t, and M. As such, algorithms designed to solve this problem are hard to reason about and come with few guarantees. Also, in practice they often suffer from senitivity to initialization. Hence, we *lift* the problem to a space of measures on Θ; this lifting allows us to apply a natural heuristic to devise a convex surrogate for problem (3.1.2).

We can encode an arbitrary, weighted collection of sources as an atomic measure μ on Θ, with mass w_i at point t_i: $\mu = \sum_{i=1}^{M} w_i \delta_{t_i}$. As a consequence of the additivity and homogeneity in our observation model, the total observation of a collection of sources encoded in the measure μ is a linear function Φ of μ:

$$\Phi\mu = \int \phi(t) d\mu(t).$$

We call Φ the *forward operator*. For atomic measures of the form $\mu = \sum_{i=1}^{n} w_i \delta_{t_i}$, this clearly agrees with (3.1.1); but it is defined for all measures on Θ.

We now introduce the sparse inverse problem as an optimization problem over the Banach space of signed measures on Θ equiped with the total variation norm. To reiterate, our goal is to recover μ_{true} from an observation

$$y = \Phi\mu_{\text{true}} + \nu$$

corrupted by the noise term, ν. Recovering the signal parameters without any prior information is, in most interesting problems, impossible; the operator Φ is almost never injective. However, in a sparse inverse problem we have the prior belief that the number of sources present, while still unknown, is small. That is, we assume that μ_{true} is an atomic measure supported on very few points.

To make the connection to compressed sensing clear, we refer to such measures as *sparse* measures. Note that while we are using the language of *recovery* or *estimation* in this section, the optimization problem we introduce is also applicable in cases where these may not be a true measure underlying the observation model. In Section 3.2 we give several examples that are not recovery problems.

We estimate the signal parameters encoded in μ_{true} by minimizing the loss ℓ of the residual between y and $\Phi\mu$:

$$\begin{array}{ll} \text{minimize} & \ell\left(\Phi\mu - y\right) \\ \text{subject to} & |\text{supp}(\mu)| \leq N. \end{array} \tag{3.1.3}$$

where the optimization is over the Banach space of signed measures (on Θ) equipped with the total variation norm. Here N is a posited upper bound on the size of the support of the true measure μ_{true}, which we denote by $\text{supp}(\mu_{\text{true}})$. Although here and elsewhere in the chapter we explicitly we place no constraint on the sign of μ, all of our discussion and algorithms can be easily extended to the nonnegative case (that is, $w \geq 0$).

While the objective function in (3.1.3) is convex, the constraint on the support of μ is nonconvex. A common heuristic in this situation is to replace the nonconvex constraint with a convex surrogate. The standard surrogate for a cardinality constraint on a measure is a constraint on the total variation. This substitution results in the standard convex approximation to (3.1.3):

$$\begin{array}{ll} \text{minimize} & \ell\left(\Phi\mu - y\right) \\ \text{subject to} & |\mu|(\Theta) \leq \tau. \end{array} \tag{3.1.4}$$

Here $\tau > 0$ is a parameter that controls the total mass of μ and empirically controls the cardinality of solutions to (3.1.4). While problem (3.1.4) is convex, it is over an infinite-dimensional space, and it is not possible to represent an arbitrary measure in a computer.

A priori, an approximate solution to (3.1.4) may have arbitrarily large support, though we prove in Section 3.5 that we can always find solutions supported on at most $d+1$ points. In practice, however, we are interested in approximate solutions of (3.1.4) supported on far fewer than $d+1$ points.

In this chapter, we propose an algorithm to solve (3.1.4) in the case where Θ has some differential structure and is therefore amenable to local, derivative based optimization. Our algorithm is based on a variant of the conditional gradient method that takes advantage of the differentiable nature of ϕ, and is guaranteed to produce approximate solutions with bounded support.

Relationship to the lasso. Readers familiar with techniques for estimating sparse vectors may recognize (3.1.4) as a continuous analogue of the standard lasso. In particular, the standard lasso is an instance of (3.1.4) with $\ell(r) = \frac{1}{2}\|r\|_2^2$ and $\Theta = \{1, \dots, k\}$. In that case, a measure over Θ can be represented as a vector v in $\mathbb{R}^k$ and the forward operator Φ as a matrix in $\mathbb{R}^{d \times k}$. The total variation of the measure v is then simply $\sum_i |v_i| = \|v\|_1$. We caution the reader that this discrete setup is substantially different as the parameter space has no differential structure. However, to make the connection to the finite dimensional case clear, we will use the notation $\|\mu\|_1$ to refer to the total varation of the measure μ.

Relationship to atomic norm problems

Problems similar to (3.1.4) have been studied through the lens of atomic norms [35]. The atomic norm $\|\cdot\|_{\mathcal{A}}$ corresponding to a suitable collection of atoms $\mathcal{A} \subset \mathbb{R}^d$ is defined as

$$\|x\|_{\mathcal{A}} = \inf \left\{ \sum_{a \in \mathcal{A}} |c_a| : x = \sum_{a \in \mathcal{A}} c_a a \right\}.$$

The connection to (3.1.4) becomes clear if we take $\mathcal{A} = \{\phi(t) : t \in \Theta\}$. With this choice of atomic set, we have the equality

$$\|x\|_{\mathcal{A}} = \inf \left\{ \|\mu\|_1 : x = \int \phi(t) d\mu(t) \right\}.$$

This equality implies the equivalence (in the sense of optimal objective value) of the infinite-dimensional optimization problem (3.1.4) to the finite-dimensional atomic norm problem:

$$\begin{array}{ll} \text{minimize} & \ell\left(x - y\right) \\ \text{subject to} & \|x\|_{\mathcal{A}} \leq \tau. \end{array} \tag{3.1.5}$$

Much of the literature on sparse inverse problems focuses on problem (3.1.5), as opposed to the infinite-dimensional problem (3.1.4). This focus is due to the fact that (3.1.5) has algorithmic and theoretical advantages over (3.1.4). First and foremost, (3.1.5) is finite-dimensional, which means that standard convex optimization algorithms may apply. Additionally, the geometry of the atomic norm ball, $\text{conv}\{\phi(t) : t \in \Theta\}$, gives clean geometric insight into when the convex heuristic will work [35].

With that said, we hold that the infinite-dimensional formulation we study has distinct practical advantages over the atomic norm problem (3.1.5). In many applications, it is the atomic decomposition that is of interest, and *not* the optimal point $x_\star$ of (3.1.5); reconstructing the optimal $\mu_\star$ for problem (3.1.4) from $x_\star$ can be highly nontrivial. For example, when designing radiation therapy, the measure $\mu_\star$ encodes the optimal beam plan directly, while the vector $x_\star = \Phi\mu_\star$ is simply the pattern of radiation that the optimal plan produces. For this reason, an algorithm that simply returns the vector $x_\star$, without the underlying atomic decomposition, is not always useful in practice.

Additionally, the measure-theoretic framework exposes the underlying parameter space, which in many applications comes with meaningful and useful structure—and is more intuitive for practitioners than the corresponding atomic norm. Naïve interpretation of the finite-dimensional optimization problem treats the parameter space as an unstructured set. Keeping the structure of the parameter space in mind makes extensions such as ADCG that make local movements in parameter space natural and uniform across applications.

3.2 Example applications

Many practical problems can be formulated as instances of (3.1.4). In this section we briefly outline a few examples to motivate our study of this problem.

Superresolution imaging. The diffraction of light imposes a physical limit on the resolution of optical images. The goal of superresolution is to remove the blur induced by diffraction as well as the effects of pixelization and noise. For images composed of a collection of point sources of light, this can be posed as a sparse inverse problem as follows. The parameters $t_1, \dots, t_M$ denote the locations of M point sources (in $\mathbb{R}^2$ or $\mathbb{R}^3$), and w_i denotes the intensity, or brightness, of the ith source. The image of the ith source is given by $w_i\phi(t_i)$, where ϕ is the pixelated point spread function of the imaging apparatus.

By solving a version of (3.1.4) it is sometimes possible to localize the point sources better than the diffraction limit—even with extreme pixelization. Astronomers use this framework to deconvolve images of stars to angular resolution below the Rayleigh limit [89]. In biology this tool has revolutionized imaging of subcellular features [46, 97]. A variant of this framework allows imaging through scattering media [76]. In Section 3.6, we show that our algorithm improves upon the current state of the art for localizing point sources in a fluorescence microscopy challenge dataset.

Linear system identification. Linear time-invariant (LTI) dynamical systems are used to model many physical systems. Such a model describes the evolution of an output $y_k \in \mathbb{R}$ based on the input $u_k \in \mathbb{R}$, where $k \in \mathbb{Z}_+$ indexes time. The internal state at time k of the system is parameterized by a vector $x_k \in \mathbb{R}^m$, and its relationship to the output is described by

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k. \end{aligned}$$

Here C is a fixed matrix, while x_0, A, and B are unknown parameters.

Linear system identification is the task of learning these unknown parameters from input-output data—that is a sequence of inputs $u_1, \ldots, u_T$ and the observed sequence of outputs $y_1, \ldots, y_T$ [101, 53]. We pose this task as a sparse inverse problem. Each source is a small LTI system with 2-dimensional state—the measurement model gives the output of the small system on the given input. To be concrete, the parameter space Θ is given by tuples of the form (x_0, r, α, B) where x_0 and B both lie in the ℓ_∞ unit ball in $\mathbb{R}^2$, r is in $[0, 1]$, and α is in $[0, \pi]$. The LTI system that each source describes has

$$A = r \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

The mapping ϕ from the parameters (x_0, r, α, B) to the output of the corresponding LTI system on input $u_1, \ldots, u_T$ is differentiable. In terms of the overall LTI system, adding the output of two weighted sources corresponds to concatenating the corresponding parameters.

In Section 3.6, we show that our algorithm matches the state of the art on two standard system identification datasets.

Matrix completion. The task of matrix completion is to estimate all entries of a large matrix given observations of a few entries. Clearly this task is impossible without prior information or assumptions about the matrix. If we believe that a low-rank matrix will approximate the truth well, a common heuristic is to minimize the squared error subject to a nuclear norm bound. For background in the theory and practice of matrix completion under this assumption see [7, 28]. We solve the following optimization problem:

$$\min_{\|A\|_* \leq \tau} \|\Gamma(A) - y\|^2.$$

Here Γ is the masking operator, that is, the linear operator that maps a matrix $A \in \mathbb{R}^{n \times m}$ to the vector containing its observed entries, and y is the vector of observed entries. We can rephrase this in our notation by letting $\Theta = \{(u, v) \in \mathbb{R}^n \times \mathbb{R}^m : \|u\|_2 = \|v\|_2 = 1\}$, $\phi((u, v)) = \Gamma(uv^T)$, and $\ell(\cdot) = \|\cdot\|^2$. In Section 3.6, we show that our algorithm achieves state of the art results on the Netflix Challenge, a standard benchmark in matrix completion.

Bayesian experimental design. In experimental design we seek to estimate a vector $x \in \mathbb{R}^d$ from measurements of the form

$$y_i = f(t_i)^T x + \epsilon_i.$$

Here $f : \Theta \to \mathbb{R}^d$ is a known differentiable feature function and ϵ_i are independent noise terms. We want to choose $t_i, \ldots, t_M$ to minimize our uncertainty about x — if each measurement requires a costly experiment, this corresponds to getting the most information from a fixed number of experiments. For background, see [88].

In general, this task in intractable. However, if we assume ϵ_i are independently distributed as standard normals and x comes from a standard normal prior we can analytically derive the posterior distribution of x given $y_1, \ldots, y_M$, as the full joint distribution of $x, y_1, \ldots, y_M$ is normal.

One notion of how much information $y_1, \ldots, y_M$ carry about x is the entropy of the posterior distribution of x given the measurements. We can then choose $t_1, \ldots, t_M$ to minimize the entropy of the posterior, which is equivalent to minimizing the (log) volume of an uncertainty ellipsoid. With this setup, the posterior entropy is (up to additive constants and a positive multiplicative factor) simply

$$-\log\det\left(I + \sum_i f(t_i)f(t_i)^T\right)^{-1}.$$

To put this in our framework, we can take $\phi(t) = f(t)f(t)^T$, $y = 0$ and $\ell(A) = -\log\det(I + A)^{-1}$. We relax the requirement to choose exactly M measurement parameters and instead search for a sparse measure with bounded total mass, giving us an instance of (3.1.4).

Fitting mixture models to data. Given a parametric distribution $P(x|t)$ we consider the task of recovering the components of a mixture model from i.i.d. samples. For background see [69]. To be more precise, we are given data $\{x_1, \ldots, x_d\}$ sampled i.i.d. from a distribution of the form $P(x) = \int_{t\in\Theta} P(x|t)\pi(t)$. The task is to recover the mixing distribution π. If we assume π is sparse, we can phrase this as a sparse inverse problem. To do so, we choose $\phi(t) = (P(x_i|t))_{i=1}^d$. A common choice for ℓ is the (negative) log-likelihood of the data: i.e., $y = 0$, $\ell(p) = -\sum_i \log p_i$. The obvious constraints here are $\int d\pi(t) \leq 1, \pi \geq 0$.

Design of numerical quadrature rules. In many numerical computing applications we require fast procedures to approximate integration against a fixed measure. One way to do this is use a quadrature rule:

$$\int f(t)dp(t) \simeq \sum_{i=1}^M w_i f(x_i).$$

The quadrature rule, given by $w_i \in \mathbb{R}$ and $t_i \in \Theta$, is chosen so that the above approximation holds for functions f in a certain function class. The pairs (w_i, t_i) are known as quadrature nodes. In practice, we want quadrature rules with very few nodes to speed evaluation of the rule.

Often we don't have an a priori description of the function class from which f is chosen, but we might have a finite number of examples of functions in the class, $f_1, \ldots, f_d$, along with their integrals against p, $y_1, \ldots, y_d$. In other words, we know that

$$\int f_i(t)dp(t) = y_i.$$

A reasonable quadrature rule should approximate the integrals of the known f_i well.

We can phrase this task as a sparse inverse problem where each source is a single quadrature node. In our notation, $\phi(t) = (f_1(t), \ldots, f_d(t))$. Assuming each function f_i is differentiable, ϕ is differentiable. A common choose of ℓ for this application is simply the squared loss. For more discussion of the design of quadrature rules using the conditional gradient method, see [8, 73].

Neural spike identification. In this example we consider the voltage v recorded by an extracellular electrode implanted in the vicinity of a population of neurons. Suppose that this population of neurons contains K types of neurons, and that when a neuron of type k fires at time $t \in \mathbb{R}$, an action potential of the form $\phi(t, k)$ is recorded. Here $\phi : \mathbb{R} \times \{1, \dots, K\} \to \mathbb{R}^d$ is a vector of voltage samples. If we denote the parameters of the ith neuron by $t_i = (t_i, k_i)$, then the total voltage $v \in \mathbb{R}^d$ can be modeled as a superposition of these action potentials:

$$v = \sum_{i=1}^{M} w_i \phi(t_i).$$

Here the weights $w_i > 0$ can encode the distance between the ith neuron and the electrode. The sparse inverse problem in this application is to recover the parameters $t_1, \dots, t_M$ and weights $w_1, \dots, w_M$ from the voltage signal v. For background see [45].

Designing radiation therapy. External radiation therapy is a common treatment for cancer in which several beams of radiation are fired at the patient to irradiate tumors. The collection of beam parameters (their intensities, positions, and angles) is called the treatment plan, and is chosen to minimize an objective function specified by an oncologist. The objective usually rewards giving large doses of radiation to tumors, and low dosages to surrounding healthy tissue and vital organs. Plans with few beams are desired as repositioning the emitter takes time—increasing the cost of the procedure and the likelihood that the patient moves enough to invalidate the plan.

A beam fired with intensity $w > 0$ and parameter t delivers a radiation dosage $w\phi(t) \in \mathbb{R}^d$. Here the output is interpreted as the radiation delivered to each of d voxels in the body of a patient. The radiation dosage from beams with parameters $t_1, \dots, t_M$ and intensities $w_1, \dots, w_M$ add linearly, and the objective function is convex. For background see [59].

3.3 Conditional gradient method

In this section we present our main algorithmic development. We begin with a review of the classical conditional gradient method (CGM) for finite-dimensional convex programs. We then apply the CGM to the sparse inverse problem (3.1.4). In particular, we augment this algorithm with an aggressive local search subroutine that significantly improves the practical performance of the CGM.

The classical CGM solves the following optimization problem:

$$\text{minimize}_{x \in \mathcal{C}} f(x), \tag{3.3.1}$$

where $\mathcal{C}$ is a closed, bounded, and convex set and f is a differentiable convex function.

CGM proceeds by iteratively solving linearized versions of (3.3.1). At iteration k, we form the standard linear approximation to the function f at the current point x_k:

$$\hat{f}_k(s) \geq f(x_k) + f'(s - x_k; x_k).$$

Here $f'(s - x_k; x_k)$ is the directional derivative of the function f at x_k in the direction $s - x_k$. As f is convex, this approximation is a global lower bound. We then minimize the linearization over the feasible set to get a potential solution s_k. As s_k minimizes a simple approximation of f that degrades with distance from x_k we take a convex combination of s_k and x_k as the next iterate. We summarize this method in Algorithm 1.

Algorithm 1 Conditional gradient method (CGM)

```
For $k = 1, \ldots k_{\max}$
    1. Linearize: $\hat{f}_k(s) \leftarrow f(x_k) + f'(s - x_k; x_k)$.
    2. Minimize: $s_k \ni \operatorname{argmin}_{s \in \mathcal{C}} \hat{f}_k(s)$.
    3. Tentative update: $\tilde{x}_{k+1} \leftarrow \frac{k}{k+2} x_k + \frac{2}{k+2} s_k$.
    4. Final update: Choose $x_{k+1}$ such that $f(x_{k+1}) \leq f(\tilde{x}_{k+1})$.
```

It is important to note that minimizing $\hat{f}_k(s)$ over the feasible set $\mathcal{C}$ in step 2 may be quite difficult and requires an application-specific subroutine.

One of the more remarkable features of the CGM is step 4. While the algorithm converges using only the tentative update in step 3, all of the convergence guarantees of the algorithm are preserved if one replaces $\tilde{x}_{k+1}$ with *any* feasible x_{k+1} that achieves a smaller value of the objective. There are thus many possible choices for the final update in step 4, and the empirical behavior of the algorithm can be quite different for different choices. One common modification is to do a line-search:

$$x_{k+1} = \underset{x \in \operatorname{conv}(x_k, s_k)}{\operatorname{argmin}} f(x).$$

We use conv to denote the convex hull—in this last example, a line segment. Another variant, the *fully-corrective* conditional gradient method, chooses

$$x_{k+1} = \underset{x \in \operatorname{conv}(x_k, s_1, \ldots, s_k)}{\operatorname{argmin}} f(x).$$

In the next section, we propose a natural choice for this step in the case of measures that uses local search to speed-up the convergence of the CGM.

One appealing aspect of the CGM is that it is very simple to compute a lower bound on the optimal value $f_\star$ as the algorithm runs. As $\hat{f}_k$ lower-bounds f, we have

$$f(s) \geq \hat{f}_k = f(x_k) + f'(s - x_k; x_k) = \hat{f}_k(s)$$

for any $s \in \mathcal{C}$. Minimizing both sides over s gives us the elementary bound

$$f_\star \geq \hat{f}_k(s_k).$$

The right hand side of this inequality is readily computed after step (2). One can prove that the bound on suboptimality derived from this inequality decreases to zero ([60]), which makes it a very useful termination condition.

CGM for sparse inverse problems

In this section we apply the classical CGM to the sparse inverse problem (3.1.4). We give two versions—first a direct translation of the fully corrective variant and then our improved algorithm that leverages local search on Θ. To make it clear that we operate over the space of measures on Θ we change notation and denote the iterate by μ_k instead of x_k. The most obvious challenge is that we cannot represent a general measure on a computer unless it is finitely-supported. We will see however that the steps of CGM can in fact be carried out on a computer in this context. Moreover we later prove that the iterates can be represented with bounded memory.

Before we describe the algorithm in detail, we first explain how to linearize the objective function and minimize the linearization. In the space of measures, linearization is most easily understood in terms of the directional derivative.

In our formulation (3.1.4), $f(\mu) = \ell(\Phi\mu_k - y)$. If we define the *residual* as $r_k = \Phi\mu_k - y$, we can compute the directional derivative of our particular choice of f at μ_k in the direction of the measure s as

$$\begin{aligned} f'(s;\mu_k) &= \lim_{c\downarrow 0} \frac{\ell(\Phi(\mu_k + cs) - y) - \ell(\Phi(\mu_k) - y)}{c} = \lim_{t\downarrow 0} \frac{\ell(r_k + c\Phi s) - \ell(r_k)}{c} = \ell'(\Phi s; r_k) \\ &= \langle \nabla\ell(r_k), \Phi s\rangle \,. \end{aligned} \tag{3.3.2}$$

Here, the inner product on the right hand side of the equation is the standard inner product in $\mathbb{R}^d$.

The second step of the CGM minimizes the linearized objective over the constraint set. In other words, we minimize $\langle \nabla\ell(r_k), \Phi s\rangle$ over a candidate measure s with total variation bounded by τ. Interchanging the integral (in Φ) with the inner product, and defining $F(t) := \langle \nabla\ell(r_k), \phi(t)\rangle$, we need to solve the optimization problem:

$$\underset{|s|(\Theta)\leq\tau}{\text{minimize}} \int F(t)ds(t). \tag{3.3.3}$$

The optimal solution of (3.3.3) is the point-mass $-\tau\text{sgn}(F(t_\star))\delta_{t_\star}$, where $t_\star \in \operatorname{argmax}|F(t)|$. This means that at each step of the CGM we need only add a single point to the support of our approximate solution μ_k. Moreover we prove that our algorithm produces iterates μ_k with support on at most $d+1$ points (see Theorem 3.5.1).

We now describe the fully-corrective variant of the CGM for sparse inverse problems (Algorithm 2). The state of the algorithm at iteration k is an atomic measure μ_k supported on a finite set S_k with mass $\mu_k(\{t\})$ on points $t \in S_k$. The algorithm alternates between selecting a source to add to the support, and tuning the weights to lower the current cost. This tuning step (Step 4) is a finite-dimensional convex optimization problem that we can solve with an off-the-shelf algorithm.

Algorithm 2 Conditional gradient method for measures (CGM-M)

```
For $k = 1 : k_{\max}$
    1. Compute gradient of loss:    $g_k = \nabla \ell(\Phi \mu_{k-1} - y)$.
    2. Compute next source:         $t_k \in \operatorname*{argmax}_{t \in \Theta} |\langle g_k, \phi(t) \rangle|$.
    3. Update support:              $S_k \leftarrow S_{k-1} \cup \{t_k\}$.
    4. Compute weights:             $\mu_k \leftarrow \operatorname*{argmin}_{\substack{|\mu|(S_k) \leq \tau \\ |\mu|(S_k^c) = 0}} \ell\left(\sum_{t \in S_k} \mu(\{t\}) \phi(t) - y\right)$.
    5. Prune support:               $S_k \leftarrow \operatorname{supp}(\mu_k)$.
```

We stress here that the objective in step 2 is *nonlinear* in the parameter t, but *linear* when considered as a functional of the measure s_k.

While we can simply run for a fixed number of iterations, we may stop early using the standard CGM bound. With a tolerance parameter $\epsilon > 0$, we terminate when the conditional gradient bound assures us that we are at most ϵ-suboptimal. In particular, we terminate when

$$\tau |\langle \phi(t_k), g_k \rangle| - \langle \Phi \mu_k, g_k \rangle < \epsilon. \tag{3.3.4}$$

Unfortunately, CGM-M does not perform well in practice. Not only does it converge very slowly, but the solution it finds is often supported on an undesirably large set. As illustrated in Figure 3.1, the performance of CGM-M is limited by the fact that it can only change the support of the measure by adding and removing points; it cannot smoothly move S_k within Θ. Figure 3.1 shows CGM-M applied to an image of two closely separated sources. The first source t_1 is placed in a central position overlapping both true sources. In subsequent iterations sources are placed too far to the right and left, away from the true sources. To move the support of the candidate measure requires CGM-M to repeatedly add and remove sources; it is clear that the ability to move the support smoothly within the parameter space would resolve this issue immediately.

In practice, we can speed up convergence and find significantly sparser solutions by allowing the support to move continuously within Θ. The following algorithm, which we call the alternating descent conditional gradient method (ADCG), exploits the differentiability of ϕ to locally improve the support at each iteration.

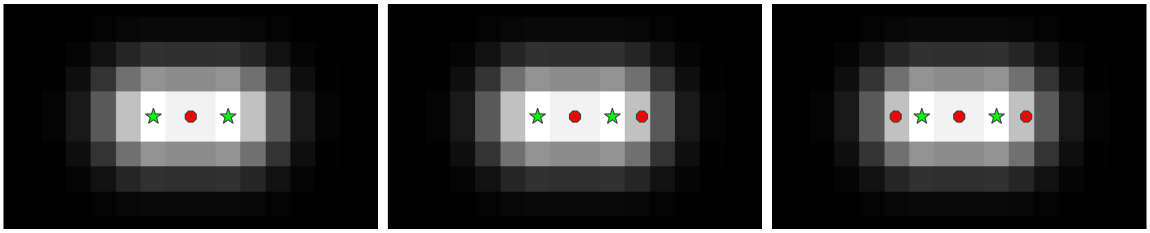

Figure 3.1: The three plots above show the first three iterates of the fully corrective CGM in a simulated superresolution imaging problem with two point sources of light. The locations of the true point sources are indicated by green stars, and the greyscale background shows the pixelated image. The elements of S_k for $k = 1, 2, 3$ are displayed by red dots.

```
Algorithm 3 Alternating descent conditional gradient method (ADCG)
For $k = 1 : k_{\max}$
    1. Compute gradient of loss:    $g_k \leftarrow \nabla \ell(\Phi \mu_{k-1} - y)$.
    2. Compute next source:         Choose $t_k \in \operatorname{argmax}_{t \in \Theta} |\langle \phi(t), g_k \rangle|$.
    3. Update support:              $S_k \leftarrow S_{k-1} \cup \{t_k\}$.
    4. Coordinate descent on nonconvex objective:
       Repeat:
           a) Compute weights:             $\mu_k \leftarrow \operatorname*{argmin}_{\substack{|\mu|(S_k) \leq \tau \\ |\mu|(S_k^c) = 0}} \ell\left(\sum_{t \in S_k} \mu(\{t\}) \phi(t) - y\right)$.
           b) Prune support:               $S_k = \text{support}(\mu_k)$.
           c) Locally improve support:     $S_k = \textbf{local\_descent}((t, \mu_k(\{t\})) : t \in S_k)$.
```

Here **local_descent** is a subroutine that takes a measure μ_k with atomic representation $(t_1, w_1), \ldots, (t_m, w_m)$ and attempts to use gradient information to reduce the function

$$(t_1, \ldots, t_m) \mapsto \ell\left(\sum_{i=1}^{m} w_i \phi(t_i) - y\right),$$

holding the weights fixed.

When the number of sources is held fixed, the optimization problem

$$\begin{array}{ll} \text{minimize} & \ell\left(\sum_{i=1}^{m} w_i \phi(t_i) - y\right) \\ \text{subject to} & t_i \in \Theta \\ & \sum_{i=1}^{m} |w_i| \leq \tau \end{array} \tag{3.3.5}$$

is nonconvex. Step 4 is then block coordinate descent over w_i and t_i. The algorithm as a whole can be interpreted as alternating between performing descent on the convex (but infinite-dimensional) problem (3.1.4) in step 2 and descent over the finite-dimensional (but nonconvex) problem (3.3.5) in step 4. The bound (3.3.4) remains valid and can be used as a termination condition.

As we have previously discussed, this nonconvex local search does not change the convergence guarantees of the CGM whatsoever. We will show in §3.5 that this is an immediate consequence of the existing theory on the CGM. However, as we will show in §3.6, the inclusion of this local search dramatically improves the performance of the CGM.

Interface and implementation

Roughly speaking, running ADCG on a concrete instance of (3.1.4) requires subroutines for two operations. We need algorithms to approximately compute:

(a) $\phi(t)$ and $\frac{d}{dt}\phi(t) \quad$ for $t \in \Theta$.

(b) $\underset{t\in\Theta}{\operatorname{argmax}} |\langle \phi(t), v\rangle|$ for arbitrary vectors $v \in \mathbb{R}^d$.

Computing (a) is usually straightforward in applications with differentiable measurement models. Computing (b) is not easy in general. However, there are many applications of interest where (b) is tractable. For example, if the parameter space Θ is low-dimensional, then the ability to compute (a) is sufficient to approximately compute (b): we can simply grid the parameter space and begin local search using the gradient of the function $t \mapsto \langle \phi(t), v\rangle$. Note that because of the local improvement step, ADCG works well even without exact minimization of (b). We prove this fact about inexact minimization in §3.5.

If the parameter space is high-dimensional, however, the feasibility of computing (b) will depend on the specific application. One example of particular interest that has been studied in the context of the CGM is matrix completion [61, 90, 54, 120]. In this case, the (b) step reduces to computing the leading singular vectors of a sparse matrix. We will show that adding local improvement to the CGM accelerates its convergence on matrix completion in the experiments.

We also note that in the special case of linear system identification, Θ is 6 dimensional, which is just large enough such that gridding is not feasible. In this case, we show that we can reduce the 6-dimensional optimization problem to a 2-dimensional problem and then again resort to gridding. We expect that in many cases of interest, such specialized solvers can be applied to solve the selection problem (b).

3.4 Related work

There has recently been a renewed interest in the conditional gradient method as a general purpose solver for constrained inverse problems [60, 54]. These methods are simpler to implement than the projected or proximal gradient methods which require solving a quadratic rather than linear optimization over the constraint set.

The idea of augmenting the classic conditional gradient method with improvement steps is not unique to our work. Indeed, it is well known that any modification of the iterate that decreases the objective function will not hurt theoretical convergence rates [60]. Moreover, Rao *et al* [90] have proposed a version of the conditional gradient method, called CoGENT, for atomic norm problems that take advantage of many common structures that arise in inverse problems. The reduction described in our theoretical analysis makes it clear that our algorithm can be seen as an instance of CoGENT specialized to the case of measures and differentiable measurement models.

The most similar proposals to ADCG come from the special case of matrix completion or nuclear-norm regularized problems. Several papers [120, 74, 54, 61] have proposed algorithms based on combinations of rank-one updates and local nonconvex optimization inspired by the well-known heuristic of [25]. While our proposal is significantly more general, ADCG essentially recovers these algorithms in the special case of nuclear-norm problems.

We note that in the context of inverse problems, there are a variety of algorithms proposed to solve the general infinite-dimensional problem (3.1.4). Tang *et al* [112] prove that this problem can be approximately solved by gridding the parameter space and solving the resulting finite dimensional problem. However, these gridding approaches are not tractable for problems with parameter spaces even of relatively modest dimension. Moreover, even when gridding is tractable, the solutions obtained are often supported on very large sets and heuristic post-processing is required to achieve reasonable performance in practice [112]. In spite of these limitations, gridding is the state of the art in many application areas including computational neuroscience [45], superresolution fluorescence microscopy [72], radar [11, 56], remote sensing [47], compressive sensing [9, 78, 42], and polynomial interpolation [91].

There have also been a handful of papers that attempt to tackle the infinite-dimensional problem without gridding. For the special case where $\ell(\cdot) = \|\cdot\|_2^2$, Bredies and Pikkarainen [24] propose an algorithm to solve the Tikhonov-regularized version of problem (3.1.4) that is very similar to Algorithm 3. They propose performing a conditional gradient step to update the support of the measure, followed by soft-thresholding to update the weights. Finally, with the weights of the measure fixed they perform discretized gradient flow over the locations of the point-masses. However, they do not solve the finite-dimensional convex problem at every iteration, which means there is no guarantee that their algorithm has bounded memory requirements. For the same reason, they are limited to one pass of gradient descent in the nonconvex phase of the algorithm. In §3.6 we show that this limitation has serious performance implications in practice.

3.5 Theoretical guarantees

In this section we present a few theoretical results. The first guarantees that we can run our algorithm with bounded memory. The second result guarantees that the algorithm converges to an optimal point and bounds the worst-case rate of convergence.

Bounded memory

As the CGM for measures adds one point to the support of the iterate per iteration, we know that the cardinality of the support of μ_k is bounded by k. For large k, then, μ_k could have large support. The following theorem guarantees that we can run our algorithm with bounded memory and in fact we need only store at most $d+1$ points, where d is the dimension of the measurements.

Theorem 3.5.1. *ADCG may be implemented to generate iterates with cardinality of support uniformly bounded by $d+1$.*

Proof. Lemma (3.5.2) allows us to conclude that the fully-corrective step ensures that the support of the measure remains bounded by $d+1$ for all iterations. □

Lemma 3.5.2. *The finite-dimensional problem*

$$\underset{\|w\|_1 \le \tau}{\text{minimize}} \;\; \ell(\sum_i w_i \phi(t_i) - y) \tag{3.5.1}$$

has an optimal solution $w_\star$ with at most $d+1$ nonzeros.

Proof. Let $u_\star$ be any optimal solution to (3.5.1). As $u_\star$ is feasible, we have that

$$v = \sum_i u_{\star i}\phi(t_i) \in \tau \text{conv}(\{\pm\phi(t_i) : i = 1, \dots, m\}).$$

In other words, $\frac{v}{\tau}$ lies in the convex hull of a set in $\mathbb{R}^d$. Caratheodory's theorem immediately tells us that $\frac{v}{\tau}$ can be represented as a convex combination of at most $d+1$ points from $\{\pm\phi(t_i) : i = 1, \dots, m\}$. That is, there exists a $w_\star$ with at most $d+1$ nonzeros such that

$$\sum_{i=1}^m w_{\star i}\phi(t_i) = v.$$

This implies that $w_\star$ is also optimal for (3.5.1). □

Note that in order to find $w_\star$, we need to either use a simplex-type algorithm to solve (3.5.1) or explore the optimal set using the random ray-shooting procedure as described in [105].

Convergence analysis

We now analyze the worst-case convergence rate for ADCG applied to (3.1.4). We note that the standard proofs for the convergence of the Frank-Wolfe algorithm [60] extend immediately to the optimization in general Banach spaces. We take a different approach here by reducing to the finite-dimensional atomic norm problem; we feel that this reduction gives additional intuition and avoids potential issues with analysing algorithms in infinite-dimensional settings.

Theorem 3.5.3 below guarantees that ADCG achieves accuracy δ in $\mathcal{O}(\frac{1}{\delta})$ iterations.

The theorem applies even when the linear minimization step is performed approximately. That is, we allow t_k to be chosen such that

$$|\langle \phi(t_k), g_k \rangle| \leq \max_{t \in \Theta} |\langle \phi(t), g_k \rangle| + \frac{\zeta}{k+2} \tag{3.5.2}$$

for some $\zeta \geq 0$. When inequality (3.5.2) holds, we say that the linear minimization problem in iteration k is solved to precision ζ.

The analysis relies on a finite-dimensional optimization problem equivalent to (3.1.4). Let $\mathcal{A} = \{\pm\phi(t) : t \in \Theta\}$. Readers familiar with the literature on atomic norms [35] will recognize the finite-dimensional problem we consider as an atomic norm problem:

$$\begin{array}{ll} \text{minimize}_{x \in \mathbb{R}^d} & \ell(x - y) \\ \text{subject to} & x \in \tau \text{conv} \mathcal{A}. \end{array} \tag{3.5.3}$$

The connection to (3.1.4) becomes clear if we note that $\tau\text{conv}\mathcal{A} = \{\Phi\mu : \|\mu\|_1 \leq \tau\}$. Any feasible measure μ for (3.1.4) gives us a feasible point $\Phi\mu$ for (3.5.3). Likewise, any feasible x for (3.5.3) can be decomposed as a feasible measure μ for (3.1.4). Furthermore, these equivalences preserve the objective value.

Before we state the theorem precisely, we introduce some notation. Let $\ell_\star = \ell(\Phi\mu_\star - y)$ denote the optimal value of (3.1.4)—the discussion above implies that $\ell_\star$ is also the optimal value of (3.5.3). Following Jaggi in [60], we define the curvature parameter $C_{f,\mathcal{S}}$ of a function f on a set $\mathcal{S}$. Intuitively, $C_{f,\mathcal{S}}$ measures the maximum divergence between f and its first-order approximations, $\hat{f}(z; x) = f(x) + \langle z - x, \nabla f(x) \rangle$:

$$C_{f,\mathcal{S}} = \sup_{\substack{x,s \in \mathcal{S} \\ \gamma \in [0,1] \\ z = x + \gamma(s - x)}} \frac{2}{\gamma^2}(f(z) - \hat{f}(z; x)).$$

Theorem 3.5.3. *Let C be the curvature parameter of the function $f(x) = \ell(x - y)$ on the set $\tau\text{conv}\mathcal{A}$. If each linear minimization subproblem is solved to precision $C\zeta$, the iterates $\mu_1, \mu_2, \ldots$ of ADCG applied to (3.1.4) satisfy*

$$\ell(\Phi\mu_k - y) - \ell_\star \leq \frac{2C}{k+2}(1 + \zeta).$$

Proof. We first show that the points $\Phi\mu_1, \Phi\mu_2, \ldots$ are iterates of the standard CGM (with a particular choice of the final update step) applied to the finite-dimensional problem (3.5.3). We then appeal to [60] to complete the proof.

Suppose that $\Phi\mu_k = x_k$. We show that the linearization step in both algorithms produces the same result (up to the equivalence mentioned earlier). Let

$$t_{k+1} = \underset{t \in \Theta}{\text{argmax}} \, |\langle \phi(t), \nabla\ell(\Phi\mu_k - y) \rangle|$$

be the output of step 2 of ADCG. Let s_k be the output of the linear minimization step of the standard CGM applied to (3.5.3) starting at x_k. Then

$$s_k = \underset{s \in \tau\text{conv}\mathcal{A}}{\text{argmin}} \langle s, \nabla\ell(x_k - y) \rangle.$$

Recalling that $\text{conv}\mathcal{A} = \{\Phi\mu \mid \|\mu\|_1 \leq 1\}$, we must have $s_k = \pm\tau\phi(t_k)$. Therefore, the linear minimization steps of the standard CGM and ADCG coincide.

We now need to show that the nonconvex coordinate descent step in ADCG is a valid final update step for the standard CGM applied to (3.5.3). This is clear as the coordinate descent step does at least as well as the fully-corrective step. We can hence appeal to the results of Jaggi [60] that bound the convergence rate of the standard CGM on finite-dimensional problems to finish the proof. □

3.6 Numerical results

In this section we apply ADCG to three of the examples in §3.2: superresolution fluorescence microscopy, matrix completion, and system identification. We have made a simple implementation of ADCG publicly available on github:

https://github.com/nboyd/SparseInverseProblems.jl.

This allows the interested reader to follow along with these examples, and, hopefully, to apply ADCG to other instances of (3.1.4).

For each example we briefly describe how we implement the required subroutines for ADCG, though again the interested reader may want to consult our code for the full picture. We then describe how ADCG compares to prior art. Finally, we show how ADCG improves on the standard fully-corrective conditional gradient method for measures (CGM-M) and a variant of the gradient flow algorithm (GF) proposed in [24]. While the gradient flow algorithm proposed in [24] does not solve the finite-dimensional convex problem at each step, our version of GF does. We feel that this is a fair comparison: intuitively, fully solving the convex problem can only improve the performance of the GF algorithm. All three experiments require a subroutine to solve the finite-dimensional convex optimization problem over the weights. For this we use a simple implementation of a primal-dual interior point method, which we include in our code package.

For each experiment we select the parameter τ by inspection. For matrix completion and linear system ID this means using a validation set. For single molecule imaging each image requires a different value of τ. For this problem, we run ADCG with a large value of τ and stop when the decrease in the objective function gained by the addition of a source falls below a threshold. This heuristic can be viewed as post-hoc selection of τ and the stopping tolerance ϵ, or as a stagewise algorithm [115].

The experiments are run on a standard c4.8xlarge EC2 instance. Our naive implementations are meant to demonstrate that ADCG is easy to implement in practice and finds high-quality solutions to (3.1.4). For this reason we do not include detailed timing information.

Superresolution fluorescence microscopy

We analyze data from the Single Molecule Localization Microscopy (SMLM) challenge [98, 20]. Fluorescence microscopy is an imaging technique used in the biological sciences to study subcellular structures in vivo. The task is to recover the 2D positions of a collection of fluorescent proteins from images taken through an optical microscope.

Here we compare the performance of our ADCG to the gridding approach of Tang *et al.* [112], two algorithms from the microscopy community (quickPALM and center of Gaussians), and also CGM and the gradient flow (GF) algorithm proposed by [24]. The gridding approach approximately solves the continuous optimization problem (3.1.4) by discretizing the space Θ into a finite grid of candidate point source locations and running an ℓ_1-regularized regression. In practice there is typically a small cluster of nonzero weights in the neighborhood of each true point source. With a fine grid, each of these clusters contains many nonzero weights, yielding many false positives.

To remove these false positives, Tang *et al.* propose a heuristic post-processing step that involves taking the center of mass of each cluster. This post-processing step is hard to understand theoretically, and does not perform well with a high-density of fluorophores.

Implementation details

For this application, the minimization required in step 2 of ADCG is not difficult: the parameter space is two-dimensional. Coarse gridding followed by a local optimization method works well in theory and practice.

For **local_descent** we use a standard constrained gradient method provided by the NLopt library [64].

Evaluation

We measure localization accuracy by computing the F_1 score, the harmonic mean of precision and recall, at varying radii. Computing the precision and recall involves first matching estimated point sources to true point sources—a difficult task. Fortunately, the SMLM challenge website [20] provides a stand-alone application that we use to compute the F_1 score.

We use a dataset of 12000 images that overlay to form simulated microtubules (see Figure 3.2) available online at the SMLM challenge website [20]. There are 81049 point sources in total, roughly evenly distributed across the images. Figure 3.2a shows a typical image. Each image covers an area 6400nm across, meaning each pixel is roughly 100nm by 100nm.

Figure 3.3 compares the performance of ADCG, gridding, quickPALM, and center of Gaussians (CoG) on this dataset. We match the performance of the gridding algorithm from [112], and significantly beat both quickPALM and CoG. Our algorithm analyses all images in well under an hour—significantly faster than the gridding approach of [112]. Note that the gridding algorithm of [112] does not work without a post-processing step.

Matrix completion

As described in §3.2, matrix completion is the task of estimating an approximately low rank matrix from some of its entries. We test our proposed algorithm on the Netflix Prize dataset, a standard benchmark for matrix completion algorithms.

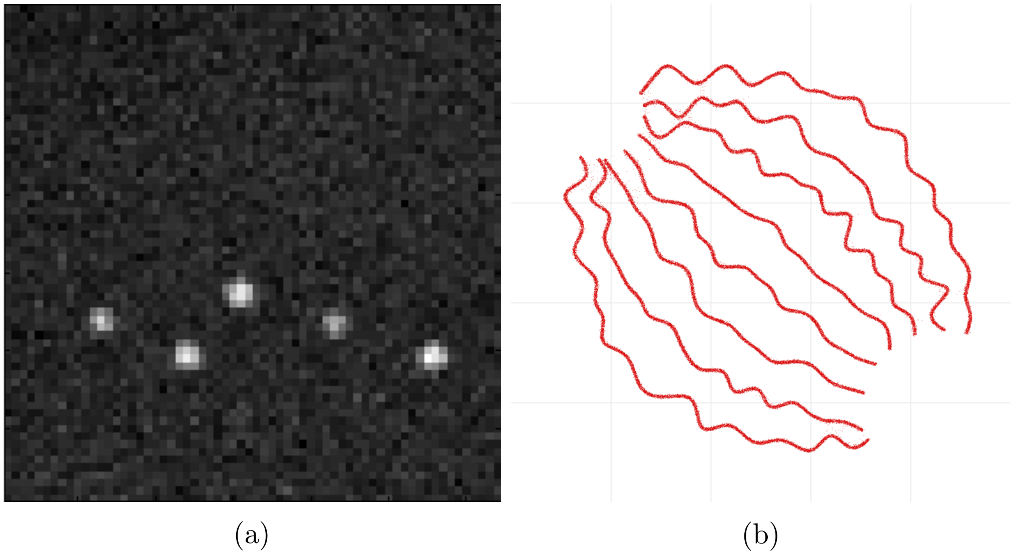

Figure 3.2: The long sequence dataset contains 12000 images similar to (a). The recovered locations for all the images are displayed in (b).

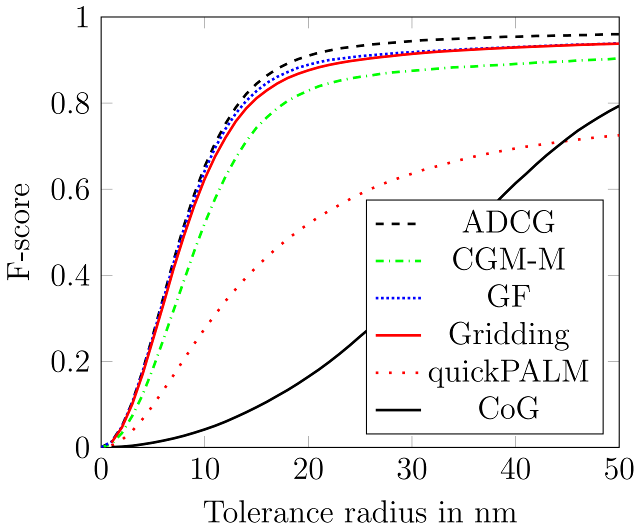

Figure 3.3: **Performance on bundled tubes: long sequence.** F-scores at various radii for 6 algorithms. For reference, each image is 6400nm across, meaning each pixel has a width of 100nm. ADCG outperforms all competing methods on this dataset.

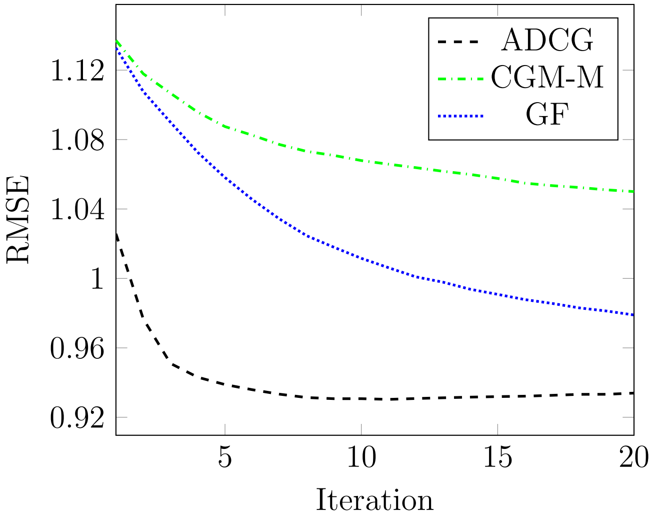

Figure 3.4: **RMSE on Netflix challenge dataset.** ADCG significantly outperforms CGM-M.

Implementation details

Although the parameter space for this example is high-dimensional we can still compute the steepest descent step over the space of measures. The optimization problem we need to solve is to minimize over all $a, b \in \mathbb{R}^n$ with $\|a\|_2 = \|b\|_2 = 1$

$$\langle \phi(a,b), \nu \rangle = \langle \Gamma(ab^T), \nu \rangle = \langle ab^T, \Gamma^*(\nu) \rangle = a^T \Gamma^*(\nu) b.$$

In other words, we need to find the unit norm, rank one matrix with highest inner product with the matrix $\Gamma^*\nu$. The solution to this problem is given by the top singular vectors of $\Gamma^*\nu$. Computing the top singular vectors using a Lanczos method is relatively easy as the matrix $\Gamma^*\nu$ is extremely sparse.

Our implementation of **local_descent** takes a single step of gradient descent (on the sphere) with line-search.

Evaluation

Our algorithm matches the state of the art for nuclear norm based approaches on the Netflix Prize dataset. Briefly, the task here is to predict the ratings 480,189 Netflix users give to a subset of 17,770 movies. One approach has been to phrase this as a matrix completion problem. That is, to try to complete the 480,189 by 17,770 matrix of ratings from the observed entires. Following [93] we subtract the mean training score from all movies and truncate the predictions of our model to lie between 1 and 5.

Figure 3.4 shows root-mean-square error (RMSE) of our algorithm and other variants of the CGM on the Netflix probe set. Again, ADCG outperforms all other CGM variants. Our algorithm takes over 7 hours to achieve the best RMSE—this could be improved with a more sophisticated implementation, or parallelization.

Comparison to prior approaches

Many researches have proposed solving matrix completion problems or general semi-definite programs using CGM-like algorithms; see [120, 74, 54, 61]. While ADCG applied to the matrix completion problem is distinct (to the best of our knowledge) from existing algorithms, it combines existing ideas. For instance, the idea of using the conditional gradient algorithm to solve the constrained formulation is very well known (see [61]). The idea of using local search on a low-rank factorization goes back at least to [25], and is used in many recent algorithms [120, 74].

In terms of performance, our implementation is relatively slow but gives very good performance in terms of validation error.

System identification

In this section we apply our algorithms to identifying two single-input single-output systems from the DaISy collection [80]: the flexible robot arm dataset (ID 96.009) and the hairdryer dataset (ID 96.006).

Implementation details

While the parameter space is 6-dimensional, which effectively precludes gridding, we can efficiently solve the minimization problem in step (2) of the ADCG. To do this, we grid only over r and α: the output is linear in the remaining parameters (B and x_0) allowing us to analytically solve for the optimal B and x_0 as a function of r and α.

For **local_descent** we again use a standard box-constrained gradient method provided by the NLopt library [64].

Evaluation

Both datasets were generated by driving the system with a specific input and recording the output. The total number of samples is 1000 in both cases. Following [101] we identify the system using the first 300 time points and we evaluate performance by running the identified system forward for the remaining time points and compare our predictions to the ground truth.

We evaluate our predictions y_{pred} using the score defined in [53]. The score is given by

$$\text{score} = 100\left(1 - \frac{\|y_{\text{pred}} - y\|_2}{\|y_{\text{mean}} - y\|_2}\right), \tag{3.6.1}$$

where y_{mean} is the mean of the test set y.

Figure 3.5 shows the score versus the number of sources as we run our algorithm. For reference we display with horizontal lines the results of [53]. ADCG matches the performance of [53] and exceeds that of all other CGM variants. Our simple implementation takes about an hour, which compares very poorly with the spectral methods in [53] which complete in under a minute.

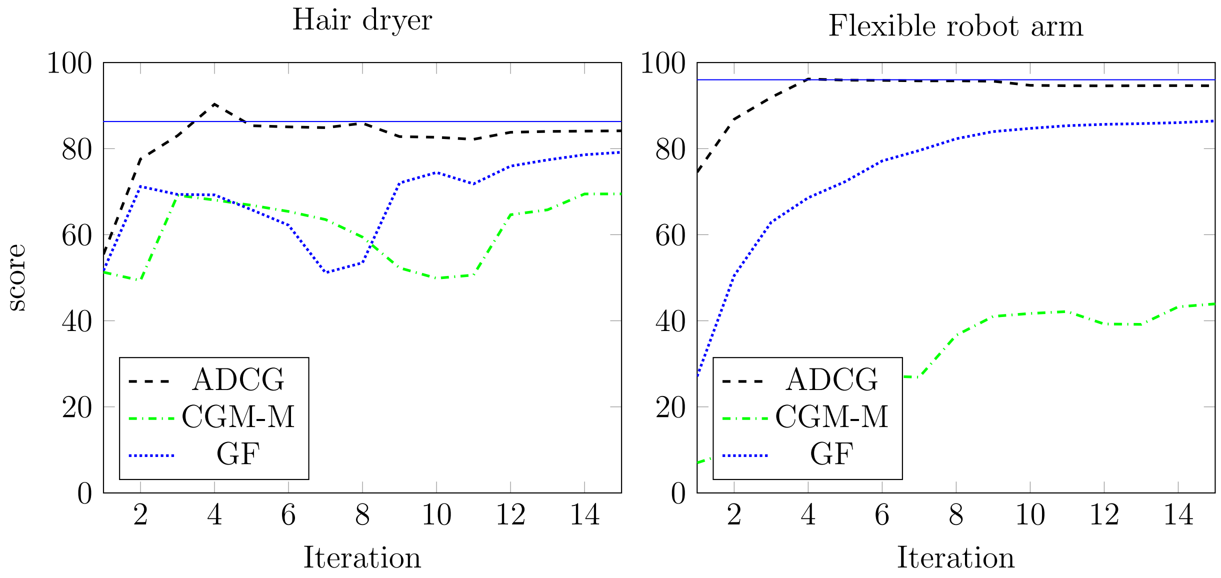

Figure 3.5: **Performance on DaISy datasets.** ADCG outperforms other CGM variants and matches the nuclear-norm based technique of [101].

3.7 Conclusions and future work

As demonstrated in the numerical experiments of Section 3.6, ADCG achieves state of the art performance in superresolution fluorescence microscopy, matrix completion, and system identification, without the need for heuristic post-processing steps. The addition of the nonconvex local search step **local_descent** significantly improves performance relative to the standard conditional gradient algorithm in all of the applications investigated. In some sense, we can understand ADCG as a method to rigorously control local search. One could just start with a model expansion (3.1.1) and perform nonconvex local search. However, this fares far worse than ADCG in practice and has no theoretical guarantees. The ADCG framework provides a clean way to generate a globally convergent algorithm that is practically efficient. Understanding this coupling between local search heuristics and convex optimization leads our brief discussion of future work.

Tighten convergence analysis for ADCG. The conditional gradient method is a robust technique, and adding our auxiliary local search step does not change its convergence rate. However, in practice, the difference between the ordinary conditional gradient method, the fully corrective variants, and ADCG are striking. In many of our experiments, ADCG outperforms the other variants by appreciable margins. Yet, all of these algorithms share the same upper bound on their convergence rate. A very interesting direction of future work would be to investigate if the bounds for ADCG can be tightened at all to be more predictive of practical performance. There may be connections between our algorithm and other alternating minimization techniques popular in matrix completion [67, 62], sparse coding [1, 5], and phase retrieval [82], and perhaps the techniques from this area could be applied to our setting of sparse inverse problems.

Relaxation to clustering algorithms. Another possible connection that could be worth exploring is the connection between the CGM and clustering algorithms like k-means. Theoretical bounds have been devised for initialization schemes for clustering algorithms that resemble the first step of CGM [6, 85]. In these methods, k-means is initialized by randomly seeking the points that are farthest from the current centers. This is akin to the first step of CGM which seeks the model parameters that best describe the residual error. Once a good seeding is acquired, the standard Lloyd iteration for k-means can be shown to converge to the global optimal solution [85]. It is possible these analyses could be generalized to analyze our version of CGM or inspire new variants of the CGM.

Connections to cutting plane methods and semi-infinite programs. The standard Lagrangian dual of (3.1.4) is a semi-infinite program (SIP), namely an optimization problem with a finite dimensional decision variable but an infinite collection of constraints [58, 102]. One of the most popular algorithmic techniques for SIP is the cutting plane method, and these methods qualitatively act very much like the CGM. Exploring this connection in detail could generate variants of cutting plane methods suited for continuous constraint spaces. Such algorithms could be valuable tools for solving semi-infinite programs that arise in contexts disjoint from sparse inverse problems.

Other applications. We believe that our techniques are broadly applicable to other sparse inverse problems, and hope that future work will explore the usefulness of ADCG in areas unexplored in this chapter. To facilitate the application of ADCG to more problems, such as those described in Section 3.2, we have made our code publicly available on GitHub. As described in Section 3.3, implementing ADCG for a new application essentially requires only two user-specified subroutines: one routine that evaluates the observation model and its derivatives at a specified set of weights and model parameters, and one that approximately solves the linear minimization in step 2 of ADCG. We aim to investigate several additional applications in the near future to test the breadth of the efficacy of ADCG.

Chapter 4

The Geometry of Kernelized Spectral Clustering

Clustering of data sets is a standard problem in many areas of science and engineering. The method of spectral clustering is based on embedding the data set using a kernel function, and using the top eigenvectors of the normalized Laplacian to recover the connected components. We study the performance of spectral clustering in recovering the latent labels of i.i.d. samples from a finite mixture of nonparametric distributions. The difficulty of this label recovery problem depends on the overlap between mixture components and how easily a mixture component is divided into two nonoverlapping components. When the overlap is small compared to the indivisibility of the mixture components, the principal eigenspace of the populationlevel normalized Laplacian operator is approximately spanned by the square-root kernelized component densities. In the finite sample setting, and under the same assumption, embedded samples from different components are approximately orthogonal with high probability when the sample size is large. As a corollary we control the fraction of samples mislabeled by spectral clustering under finite mixtures with nonparametric components.

This chapter is joint work with Martin Wainwright and Bin Yu. The content of this chapter has been published in The Annals of Statistics under the title *The Geometry of Kernelized Spectral Clustering.*

4.1 Introduction

In the past decade, spectral methods have emerged as a powerful collection of nonparametric tools for unsupervised learning, or clustering. How can we recover information about the geometry or topology of a distribution from its samples? Clustering algorithms attempt to answer the most basic form of this question. One way in which to understand spectral clustering is as a relaxation of the NP-hard problem of searching for the best graph-cut. Spectral graph partitioning—using the eigenvectors of a matrix to find graph cuts—originated in the early 1970's with the work of Fiedler [49] and of Donath and Hoffman [37]. Spectral clustering was introduced in machine learning, with applications to clustering data sets and computing image segmentations (e.g., [103, 83, 79]). The past decade has witnessed an explosion of different spectral clustering algorithms. One point of variation is that some use the eigenvectors of the kernel matrix [104, 63, 37], or adjacency matrix in the graph setting, whereas others use the eigenvectors of the normalized Laplacian matrix [83, 79, 103, 49]. This division goes all the way back to the work of Donath and Hoffman, who proposed using the adjacency matrix, and of Fiedler, who proposed using the normalized Laplacian matrix.

In its modern and most popular form, the spectral clustering algorithm [83, 103] involves two steps: first, the eigenvectors of the normalized Laplacian are used to embed the dataset, and second, the M-means clustering algorithm is applied to the embedded dataset. The normalized Laplacian embedding is an attractive preprocessing step because the transformed clusters tend to be linearly separable. Ng et al. [83] show that, under certain conditions on the empirical kernel matrix, an embedded dataset will cluster tightly around well separated points on the unit sphere. Their results apply to a fixed dataset, and do not model the underlying distribution of the data. Recently Yan et al. [119] derive an expression for the fraction of data misclustered by spectral clustering by computing an analytical expression for the second eigenvector of the Laplacian. They assume that the similarity matrix is a small perturbation away from the ideal block diagonal case.

The embedding defined by the normalized Laplacian has also been studied in the context of manifold learning, where the primary focus has been convergence of the underlying eigenvectors. This work is motivated in part by the fact that spectral properties of the limiting Laplace-Beltrami operator have long been known to shed light on the connectivity of a manifold [75]. The Laplacian eigenmaps of Belkin and Niyogi [15] reconstruct Laplace-Beltrami eigenfunctions from sampled data. Koltchinskii and Giné [70] analyze the convergence of the empirical graph Laplacian to the Laplace-Beltrami operator at a fixed point in the manifold. von Luxburg and Belkin [77] establish consistency for the embedding in as much as the eigenvectors of the Laplacian matrix converge uniformly to the eigenfunctions of the Laplacian operator. Rosasco et al. [96] provide simpler proofs of this convergence, and in part, our work sharpens these results by removing an unnecessary smoothness assumption on the kernel function.

In this chapter, we study spectral clustering in the context of a nonparametric mixture model. The study of spectral clustering under nonparametric mixtures was initiated by Shi et al. [104]. One of their theorems characterizes the top eigenfunction of a kernel integral operator, showing that it does not change sign. One difficulty in using the eigenfunctions of a kernel integral operator to separate mixture components is that several of the top eigenfunctions may correspond to a single mixture component (e.g. one with a larger mixture weight). They proposed that eigenfunctions of the kernel integral operator that approximately do not change sign correspond to different mixture components. However, their analysis does not deal with finite datasets nor does it provide bounds on the fraction of points misclustered.

The main contribution of this chapter is an analysis of the normalized Laplacian embedding of i.i.d. samples from a finite mixture with nonparametric components. We begin by providing a novel and useful characterization of the principal eigenspace of the population-level normalized Laplacian operator: more precisely, when the mixture components are indivisible and have small overlap, the eigenspace is close to the span of the square root kernelized component densities. We then use this characterization to analyze the geometric structure of the embedding of a finite set of i.i.d. samples. Our main result is to establish a certain geometric property of nonparametric mixtures referred to as *orthogonal cone structure*. In particular, we show that when the mixture components are indivisible and have small overlap, embedded samples from different components are almost orthogonal with high probability. We then prove that this geometric structure allows M-means to correctly label most of the samples. Our proofs rely on techniques from operator perturbation theory, empirical process theory, and spectral graph theory.

The remainder of this chapter is organized as follows. In Section 4.2, we set up the problem of separating the components of a mixture distribution. We state our main results and explore some of their consequences in Section 4.3. We prove our main results in Section 4.4, deferring the proofs of several supporting lemmas to the Appendix.

Notation For a generic distribution $\mathbb{P}$ on a measurable space $\mathcal{X}$, we denote the Hilbert space of real-valued square integrable functions on $\mathcal{X}$ by $L^2(\mathbb{P})$. The $L^2(\mathbb{P})$ inner product is given by $\langle f, g\rangle_{\mathbb{P}} = \int f(x)g(x)d\mathbb{P}(x)$, and it induces the norm $\|f\|_{\mathbb{P}}$. The norm $\|f\|_\infty$ is the supremum of the function f, up to sets of measure zero, where the relevant measure is understood from context. The Hilbert–Schmidt norm of an operator $\mathbf{T} : L^2(\mathbb{P}) \to L^2(\mathbb{P})$ is $|\!|\!|\mathbf{T}|\!|\!|_{\text{HS}}$ and the operator norm is $|\!|\!|\mathbf{T}|\!|\!|_{\text{op}}$. The complement of a set B is denoted by B^c. See Appendix D for an additional list of symbols.

4.2 Background and problem set-up

We begin by introducing the family of nonparametric mixture models analyzed in this chapter, and then provide some background on kernel functions, spectral clustering and Laplacian operators.

Nonparametric mixture distributions

For some integer $M \geq 2$, let $\{\mathbb{P}_m\}_{m=1}^M$ be a collection of probability measures on a compact space $\mathcal{X}$, and let the weights $\{w_m\}_{m=1}^M$ belong to the relative interior of the probability simplex in $\mathbb{R}^M$—that is, $w_m \in (0, 1)$ for all $m = 1, \ldots, M$, and $\sum_{m=1}^M w_m = 1$. This pair specifies a *finite nonparametric mixture distribution* via the convex combination

$$\bar{\mathbb{P}} := \sum_{m=1}^{M} w_m \mathbb{P}_m. \tag{4.2.1}$$

We refer to $\{\mathbb{P}_m\}_{m=1}^M$ and $\{w_m\}_{m=1}^M$ as the mixture components and mixture weights, respectively. The family of models (4.2.1) is *nonparametric*, because the mixture components are not constrained to any particular parametric family.

A random variable $\bar{X} \sim \bar{\mathbb{P}}$ can be obtained by first drawing a categorical random variable $Z \sim \text{Categorical}(w_1, \ldots, w_M)$, and conditioned on the event $\{Z = m\}$, drawing a variable from mixture component $\mathbb{P}_m$. Consequently, given a collection of samples $\{X_i\}_{i=1}^n$ drawn i.i.d. from $\bar{\mathbb{P}}$, there is an underlying set of *latent labels* $\{Z_i\}_{i=1}^n$. Thus, in the context of a mixture distribution, the clustering problem can be formalized as recovering these latent labels based on observing only the unlabeled samples $\{X_i\}_{i=1}^n$.

Of course, this clustering problem is ill-defined whenever $\mathbb{P}_j = \mathbb{P}_k$ for some $j \neq k$. More generally, recovery of labels becomes more difficult as the overlap of any pair $\mathbb{P}_j$ and $\mathbb{P}_k$

increases, or if it is “easy” to divide any component into two non-overlapping distributions. This intuition is formalized in our definition of the overlap and indivisibility parameters in Section 4.3 to follow.

Kernels and spectral clustering

We now provide some background on spectral clustering methods, and the normalized Laplacian embedding. A kernel k associated with the space $\mathcal{X}$ is a symmetric, continuous function $k : \mathcal{X} \times \mathcal{X} \to (0, \infty)$. A kernel is said to be positive semidefinite if, for any integer $n \geq 1$ and elements $x_1, \ldots, x_n \in \mathcal{X}$, the kernel matrix $A \in \mathbb{R}^{n \times n}$ with entries $A_{ij} = k(x_i, x_j)/n$ is positive semidefinite. Throughout we consider a fixed but arbitrary positive semidefinite kernel function. In application to spectral clustering, one purpose of a kernel function is to provide a measure of the similarity between data points. A canonical example is the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|_2^2)$; it is close to 1 for vectors x and x' that are relatively close, and decays to zero for pairs that are far apart.

Let us now describe the normalized Laplacian embedding, which is a standard part of many spectral clustering routines. Given n i.i.d. samples $\{X_i\}_{i=1}^n$ from $\bar{\mathbb{P}}$, the associated kernel matrix $A \in \mathbb{R}^{n \times n}$ has entries $A_{ij} = \frac{1}{n} k(X_i, X_j)$. The normalized Laplacian matrix[1] is obtained by rescaling the kernel matrix by its row sums—namely

$$L = D^{-1/2} A D^{-1/2}, \tag{4.2.2}$$

where D is a diagonal matrix with entries $D_{ii} = \sum_{j=1}^n A_{ij}$. Since L is a symmetric matrix by construction, it has an orthonormal basis of eigenvectors, and we let $\{v_1, \ldots, v_M\}$ denote the eigenvectors corresponding to the largest M eigenvalues of L. The *normalized Laplacian embedding* is defined on the basis of these eigenvectors: it is the map $\Phi_{\mathcal{V}} : \{X_1, \ldots, X_n\} \to \mathbb{R}^M$ defined by

$$\Phi_{\mathcal{V}}(X_i) := (v_{1i}, \ldots, v_{Mi}). \tag{4.2.3}$$

A typical form of spectral clustering consists of the following two steps. First, compute the normalized Laplacian, and map each data point X_i to a M-vector via the embedding (4.2.3). The second step is to apply a standard clustering method (such as M-means clustering) to the embedded data points. The conventional rationale for the second step is that the embedding step typically helps reveal cluster structure in the data set, so that it can be found by a relatively simple algorithm. The goal of this chapter is to formalize the sense in which the normalized Laplacian embedding (4.2.3) has this desirable property.

We do so by first analyzing the population operator that underlies the normalized Laplacian matrix. It is defined by the *normalized kernel function*

$$\bar{k}(x, y) := \frac{1}{\bar{q}(x)}\, k(x, y)\, \frac{1}{\bar{q}(y)} \tag{4.2.4}$$

where $\bar{q}(y) = \sqrt{\int k(x, y) d\bar{\mathbb{P}}(x)}$. Note that this kernel function can be seen as a continuous analog of the normalized Laplacian matrix (4.2.2).

[1] To be precise, the matrix $I - L$ is actually the normalized graph Laplacian matrix. However, the eigenvectors of L are identical to those of $I - L$, and we find it simpler to work with L.

The normalized kernel function in conjunction with the mixture defines the *normalized Laplacian operator* $\bar{\mathbf{T}} : L^2(\bar{\mathbb{P}}) \to L^2(\bar{\mathbb{P}})$ given by

$$(\bar{\mathbf{T}}f)(\cdot) := \int \bar{k}(\cdot, y) f(y) d\bar{\mathbb{P}}(y). \tag{4.2.5}$$

Under suitable regularity conditions (see Appendix C.1 for details), this operator has an orthonormal set of eigenfunctions—with eigenvalues in $[0, 1]$—and our main results relate these eigenfunctions to the underlying mixture components $\{\mathbb{P}_m\}_{m=1}^M$.

4.3 Analysis of the normalized Laplacian embedding

This section is devoted to the statement of our main results, and discussion of their consequences. These results involve a few parameters of the mixture distribution, including its overlap and indivisibility parameters, which are defind in Section 4.3. Our first main result (Theorem 4.3.1 in Section 4.3) characterizes the principal eigenspace of the population-level normalized Laplacian operator (4.2.5), showing that it approximately spanned by the square root kernelized densities of the mixture components, as defined in Section 4.3. Our second main result (Theorem 4.3.2 in Section 4.3) provides a quantitative description of the angular structure in the normalized Laplacian embedding of a finite sample from a mixture distribution.

Cluster similarity, coupling, and indivisibility parameters

In this section, we define some parameters associated with any nonparametric mixture distribution, as viewed through the lens of a given kernel. These quantities play an important role in our main results, as they reflect the intrinsic difficulty of the clustering problem.

Our first parameter is the *similarity index* of the mixture components $\{\mathbb{P}_m\}_{m=1}^M$. For any pair of distinct indices $\ell \neq m$, the ratio

$$\mathcal{S}(\mathbb{P}_\ell, \mathbb{P}_m) := \frac{\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) d\mathbb{P}_m(x) d\mathbb{P}_\ell(y)}{\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) d\bar{\mathbb{P}}(x) d\mathbb{P}_\ell(y)}.$$

is a kernel-dependent measure of the expected similarity between the clusters indexed by $\mathbb{P}_\ell$ and $\mathbb{P}_m$ respectively. Note that $\mathcal{S}$ is not symmetric in its arguments. The *maximum similarity* over all mixture components

$$\mathcal{S}_{\max}(\bar{\mathbb{P}}) := \max_{\substack{\ell, m = 1, \ldots, M \\ \ell \neq m}} \mathcal{S}(\mathbb{P}_\ell, \mathbb{P}_m) \tag{4.3.1}$$

measures the overlap between mixture components with respect to the kernel k.

Our second parameter, known as the coupling parameter, is defined in terms of the square root kernelized densities of the mixture components. More precisely, given any distribution $\mathbb{P}$, its *square root kernelized density* is the function $q \in L^2(\mathbb{P})$ given by

$$q(x) := \sqrt{\int k(x, y) d\mathbb{P}(y)}. \tag{4.3.2}$$

In particular, we denote the square root kernelized density of the mixture distribution $\bar{\mathbb{P}}$ by $\bar{q}$, and those of the mixture components $\{\mathbb{P}_m\}_{m=1}^M$ by $\{q_m\}_{m=1}^M$. In analogy with the normalized kernel function $\bar{k}$ from Equation (4.2.4), we also define a normalized kernel for each mixture component—namely

$$k_m(x,y) := \frac{k(x,y)}{q_m(x)q_m(y)} \qquad \text{for } m = 1, \ldots, M. \tag{4.3.3}$$

The *coupling parameter*

$$\mathcal{C}(\bar{\mathbb{P}}) := \max_{m=1,\ldots,M} \left\| k_m - w_m \bar{k} \right\|^2_{\mathbb{P}_m \otimes \mathbb{P}_m}, \tag{4.3.4}$$

measures the coupling of the spaces $L^2(\mathbb{P}_m)$ with respect to $\bar{\mathbf{T}}$. In particular, when $\mathcal{C}(\bar{\mathbb{P}}) = 0$, then the normalized Laplacian can be decomposed as the sum

$$\bar{\mathbf{T}} = \sum_{m=1}^{M} w_m \mathbf{T}_m, \tag{4.3.5}$$

where $(\mathbf{T}_m f)(y) = \int f(x) k_m(x,y) d\mathbb{P}_m(x)$ is the operator defined by the normalized kernel k_m. When the coupling parameter is no longer exactly zero but still small, then the decomposition (4.3.5) still holds in an approximate sense.

Our final parameter measures how easy or difficult it is to "split" any given mixture component $\mathbb{P}_m$ into two or more parts. If this splitting can be done easily for any component, then the mixture distribution will be hard to identify, since there is an ambiguity as to whether $\mathbb{P}_m$ defines one component, or multiple components. In order to formalize this intuition, for a distribution $\mathbb{P}$ and for a measurable subset $S \subset \mathcal{X}$, we introduce the shorthand notation $p(S) = \int_S \int_{\mathcal{X}} k(x,y) d\mathbb{P}(x) d\mathbb{P}(y)$. With this notation, the indivisibility of $\mathbb{P}$ is

$$\Gamma(\mathbb{P}) := \inf_S \frac{p(\mathcal{X}) \int_S \int_{S^c} k(x,y) d\mathbb{P}(x) d\mathbb{P}(y)}{p(S) p(S^c)}, \tag{4.3.6}$$

where the infimum is taken over all measurable subsets S such that $p(S) \in (0,1)$. The *indivisibility parameter* $\Gamma_{\min}(\bar{\mathbb{P}})$ of a mixture distribution $\bar{\mathbb{P}}$ is the minimum indivisibility of its mixture components

$$\Gamma_{\min}(\bar{\mathbb{P}}) := \min_{m=1,\ldots,M} \Gamma(\mathbb{P}_m). \tag{4.3.7}$$

Our results in the next section apply when the similarity $\mathcal{S}_{\max}(\bar{\mathbb{P}})$ and coupling $\mathcal{C}(\bar{\mathbb{P}})$ are small compared to the indivisibility $\Gamma_{\min}(\bar{\mathbb{P}})$. Some examples help illustrate when this is the case.

Example 1. Consider the one-dimensional triangular density function

$$g_{\mathbb{T}_\mu}(x) := \begin{cases} x - \mu + 1, & \text{if } x \in (\mu - 1, \mu); \\ -x + \mu + 1, & \text{if } x \in (\mu, \mu + 1); \\ 0, & \text{otherwise.} \end{cases}$$

with location $\mu > 0$. We denote corresponding distribution by $\mathbb{T}_\mu$. In this example we calculate the similarity, coupling, and indivisibility parameters for the mixture of triangular distributions $\bar{\mathbb{T}} := \frac{1}{2}\mathbb{T}_0 + \frac{1}{2}\mathbb{T}_\mu$ and the uniform kernel $k_\nu(x, y) = \frac{1}{2\nu}\mathbf{1}\{|x - y| \leq \nu\}$ with bandwidth $\nu \in (0, 1)$.[2]

Similarity It is straightforward to calculate the similarity parameter $\mathcal{S}_{\text{max}}(\bar{\mathbb{T}})$ by solving a few simple integrals. We find that

$$\mathcal{S}_{\text{max}}(\bar{\mathbb{T}}) = \frac{2(2 + \nu - \mu)_+^4}{\nu(16 - 8\nu^2 + 3\nu^3) + (2 + \nu - \mu)_+^4}.$$

Coupling To compute the coupling parameter, we must compute the kernelized densities of $\mathbb{T}_0$ and $\mathbb{T}_\mu$, and the normalized kernel functions $k_1(x, y)$ and $\bar{k}(x, y)$. Some calculation yields the following equation for the kernelized density

$$q_1^2(x) = \begin{cases} \frac{(1+\nu-x)^2}{4\nu}, & \text{if } x \in (-1-\nu, -1+\nu) \\ 1 - \frac{\nu}{2} - \frac{x^2}{2\nu}, & \text{if } x \in (-\nu, \nu) \\ \frac{(1+\nu+x)^2}{4\nu}, & \text{if } x \in (1-\nu, 1+\nu) \\ g_{\mathbb{T}_0}(x), & \text{otherwise.} \end{cases} \tag{4.3.8}$$

As can be seen in Figure 4.1, the kernelized density $q_1^2(x)$ of $\mathbb{T}_0$ is a smoothed version of $g_{\mathbb{T}_0}(x)$ that interpolates quadratically around the nondifferentiable points of $g_{\mathbb{T}_0}(x)$.

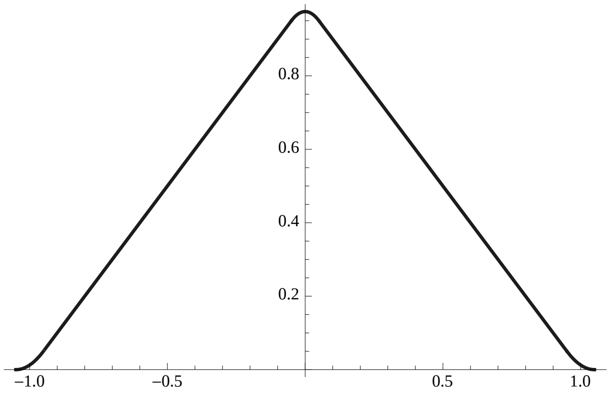

Figure 4.1: The kernelized density $q_1^2(x)$ of Equation (4.3.8) with $\nu = 0.05$.

The kernelized density of $\mathbb{T}_\mu$ has the same shape as that of $\mathbb{T}_0$ but is shifted by μ. In particular,

$$q_2^2(x) = q_1^2(x - \mu).$$

Therefore the normalized kernels satisfy $k_1(x, y) = \frac{1}{2}\bar{k}(x, y)$ for $x, y \in (-1, \mu - 1 - \nu)$. By upper bounding the integrand over the remaining region, we find that

$$\mathcal{C}(\bar{\mathbb{T}})^2 = \iint (k_1 - \bar{k}/2)^2 d\mathbb{P}_1(x, y) \leq 2(2 + \nu - \mu)_+.$$

[2] This is not a positive semidefinite kernel function, but it helps to build intuition our intuition in a case where all the integrals are easy.

Indivisibility It is straightforward to calculate the indivisibility $\Gamma(\mathbb{T}_\mu)$. For any $\nu \in (0,1)$ and $\mu \in \mathbb{R}$, the set S defining $\Gamma(\mathbb{T}_\mu)$ is $S = (\mu, \infty)$. Hence, by solving a few simple integrals we find that

$$\Gamma(\mathbb{T}_\mu) = \frac{2(6-\nu)(2-\nu)\nu}{16 - 8\nu^2 + 3\nu^3}.$$

Note that $\Gamma(\mathbb{T}_\mu)$ does not depend on μ. Therefore the indivisibility of $\bar{\mathbb{T}} := \frac{1}{2}\mathbb{T}_0 + \frac{1}{2}\mathbb{T}_\mu$ is

$$\Gamma_{\text{min}}(\bar{\mathbb{T}}) = \Gamma(\mathbb{T}_\mu) = \Gamma(\mathbb{T}_0).$$

It is instructive to consider the indivisibility of the following poorly defined two-component mixture

$$\bar{\mathbb{T}}_{\text{bad}} := \frac{1}{2}\mathbb{P}_{\text{bad}}(\mu) + \frac{1}{2}\mathbb{T}_{2\mu},$$

where $\mathbb{P}_{\text{bad}}(\mu)$ is the bimodal component $\mathbb{P}_{\text{bad}}(\mu) := \frac{1}{2}\mathbb{T}_0 + \frac{1}{2}\mathbb{T}_\mu$. It is easy to verify that for any $\nu \in (0,1)$ and $\mu > 2+\nu$, the indivisibility of the bimodal component is $\Gamma(\mathbb{P}_{\text{bad}}(\mu)) = 0$, and therefore $\Gamma_{\text{min}}(\bar{\mathbb{T}}_{\text{bad}}) = 0$.

♠

From Example 1 we learn that the similarity $\mathcal{S}_{\text{max}}(\bar{\mathbb{T}})$ and coupling $\mathcal{C}(\bar{\mathbb{T}})$ parameters decrease as the offset μ increases. Together, these two parameters measure the overlap which our intuition tells us should decrease as μ increases. On the other hand, the indivisibility parameter $\Gamma_{\text{min}}(\bar{\mathbb{T}})$ is independent of μ.

Our next example is more realistic in the sense that the kernel and mixture components do not have bounded support, and the kernel function is positive semidefinite.

Example 2. In this example we calculate the similarity, coupling and indivisibility parameters for the mixture of Gaussians $\bar{\mathbb{N}} = \frac{1}{2}\mathbb{N}(0,1) + \frac{1}{2}\mathbb{N}(\mu,1)$ equipped with the Gaussian kernel $k_\nu(x,y) = \frac{1}{\sqrt{2\pi}\nu}\exp\big[-\frac{|x-y|^2}{2\nu^2}\big]$.

Similarity As in the previous example, it is straightforward to calculate the maximal intercluster similarity $\mathcal{S}_{\text{max}}(\bar{\mathbb{N}})$ by solving a handful of Gaussian integrals. We find that

$$\mathcal{S}_{\text{max}}(\bar{\mathbb{N}}) = \frac{2\exp\frac{-\mu^2}{2\nu^2+4}}{1+\exp\frac{-\mu^2}{2\nu^2+4}} \leq 4e^{\frac{-\mu^2}{2\nu^2+4}}. \tag{4.3.9}$$

Coupling The kernelized density of $\mathbb{N}(0,1)$ is

$$q_1^2(x) = \frac{1}{\sqrt{\nu^2+1}}\exp\big[\frac{-x^2}{2(\nu^2+1)}\big],$$

and the kernelized density of $\mathbb{N}(\mu,1)$ is simply the translation $q_2^2(x) = q_1^2(x-\mu)$. We can bound the coupling parameter $\mathcal{C}(\bar{\mathbb{N}})$ by upper bounding the integrand $k_1 - \bar{k}$ over a high-probability compact set (a modification of the trick from Example 1). We show the resulting bound in Figure 4.2. This (albeit loose) bound captures the exponential decay of $\mathcal{C}(\bar{\mathbb{N}})$ with μ.

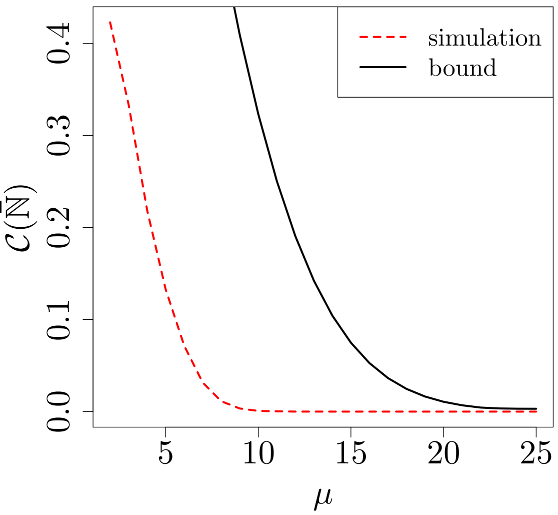

Figure 4.2. The coupling parameter for the mixture of Gaussians with Gaussian kernel and bandwidth $\nu = 2$. The red line displays the simulated value as a function of the offset μ between mixture components. The black line displays our analytical bound.

Indivisibility It is straightforward to compute the indivisibility of the unit-variance normal distribution $\mathbb{N}(\mu, 1)$ with location $\mu \in \mathbb{R}$. The set defining the indivisibility is $S = (\mu, \infty)$. Solving a handful of Gaussian integrals yields

$$\Gamma_{\min}(\mathbb{N}(\mu, 1)) = \frac{2}{\pi} \arctan(\nu\sqrt{2 + \nu^2}). \tag{4.3.10}$$

♠

We conclude with a counter example, showing that the similarity parameter is *not* relatively small for the linear kernel $k(x, y) = x \cdot y$.

Example 3. Consider the mixture $\bar{\mathbb{P}} = \frac{1}{2}\mathbb{P}_1 + \frac{1}{2}\mathbb{P}_2$ with components $\mathbb{P}_1$ uniform over $(1, 2)$ and $\mathbb{P}_2$ uniform over $(2 + \delta, 3 + \delta)$. With the linear kernel $k(x, y) = x \cdot y$, we have

$$\mathcal{S}_{\max}(\bar{\mathbb{P}}) = \frac{\iint xyd\mathbb{P}_1(x)d\mathbb{P}_2(y)}{\iint xyd\mathbb{P}_1(x)d\bar{\mathbb{P}}(y)} = \frac{\int yd\mathbb{P}_2(y)}{\int yd\bar{\mathbb{P}}(y)} \geq \frac{1}{2}.$$

Since $\Gamma_{\min}(\bar{\mathbb{P}})$ is always between 0 and 1, this calculation demonstrates that the similarity parameter $\mathcal{S}_{\max}(\bar{\mathbb{P}})$ is never small compared to $\Gamma_{\min}(\bar{\mathbb{P}})$. ♠

Population-level analysis

In this section, we present our population-level analysis of the normalized Laplacian embedding. Consider the following two subspaces of $L^2(\bar{\mathbb{P}})$:

- the subspace $\mathcal{R} \subset L^2(\bar{\mathbb{P}})$ spanned by the top M eigenfunctions of the normalized Laplacian operator $\bar{\mathbf{T}}$ from Equation (4.2.5), and
- the span $\mathcal{Q} = \text{span}\{q_1, \ldots, q_M\} \subset L^2(\bar{\mathbb{P}})$ of the square root kernelized densities (see Equation (4.3.2)).

The subspace $\mathcal{Q}$ can be used to define a map $\Phi_{\mathcal{Q}} : \mathcal{X} \to \mathbb{R}^M$ known as the *square-root kernelized density embedding*, given by

$$\Phi_{\mathcal{Q}}(x) := (q_1(x), \ldots, q_M(x)). \tag{4.3.11}$$

This map is relevant to clustering, since the vector $\Phi_{\mathcal{Q}}(x)$ encodes sufficient information to perform a likelihood ratio test (based on the kernelized densities) for labeling data points.

On the other hand, the subspace $\mathcal{R}$ is important, because it is the population-level quantity that underlies spectral clustering. As described in Section 4.2, the first step of spectral clustering involves embedding the data using the eigenvectors of the Laplacian matrix. This procedure can be understood as a way of estimating the *population-level Laplacian embedding*: more precisely, the map $\Phi_{\mathcal{R}} : \mathcal{X} \to \mathbb{R}^M$ given by

$$\Phi_{\mathcal{R}}(x) := (r_1(x), \ldots, r_M(x)), \tag{4.3.12}$$

where $\{r_m\}_{m=1}^M$ are the top M eigenfunctions of the kernel operator $\bar{\mathbf{T}}$.

To build intuition, imagine for the moment varying the kernel function so that the kernelized densities converge to the true densities. For example, imagine sending the bandwidth of a Gaussian kernel to 0. While the kernelized densities approach the true densities, the subspace $\mathcal{R}$ is only a well defined mathematical object for kernels with non-zero bandwidth. Indeed, as the bandwidth shrinks to zero, the eigengap separating the principal eigenspace $\mathcal{R}$ of $\bar{\mathbf{T}}$ from its lower eigenspaces vanishes. For this reason, we analyze an arbitrary but fixed kernel function, and we discuss kernel selection in Section 4.5.

The goal of this section is to quantify the difference between the two mappings $\Phi_{\mathcal{Q}}$ and $\Phi_{\mathcal{R}}$, or equivalently between the underlying subspaces $\mathcal{Q}$ and $\mathcal{R}$. We assume that that the square root kernelized densities $q_1, \ldots, q_M$ are linearly independent so that $\mathcal{Q}$ has the same dimension, M, as $\mathcal{R}$. This condition is very mild when the overlap parameters $\mathcal{S}_{\text{max}}(\bar{\mathbb{P}})$ and $\mathcal{C}(\bar{\mathbb{P}})$ are small. We measure the distance between these subspaces by the Hilbert–Schmidt norm[3] applied to the difference between their orthogonal projection operators:

$$\rho(\mathcal{Q}, \mathcal{R}) := |||\Pi_{\mathcal{Q}} - \Pi_{\mathcal{R}}|||_{\text{HS}}. \tag{4.3.13}$$

Recall the similarity parameter $\mathcal{S}_{\text{max}}(\bar{\mathbb{P}})$, coupling parameter $\mathcal{C}(\bar{\mathbb{P}})$, and indivisibility parameter $\Gamma_{\text{min}}(\bar{\mathbb{P}})$, as previously defined in equations (4.3.1), (4.3.4) and (4.3.7), respectively. Our main results involve a function of these three parameters and the minimum $w_{\text{min}} := \min_{m=1,\ldots,M} w_m$ of the mixture weights, given by

$$\varphi(\bar{\mathbb{P}}; k) := \frac{\sqrt{M}[\mathcal{S}_{\text{max}}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})]^{1/2}}{w_{\text{min}}\, \Gamma^2_{\text{min}}(\bar{\mathbb{P}})}. \tag{4.3.14}$$

[3] Recall that the Hilbert–Schmidt norm of an operator is the infinite dimensional analogue of the Frobenius norm of a matrix.

Our first main theorem guarantees that as long as the mixture is relatively well-separated, as measured by the *difficulty function* φ, then the ρ-distance (4.3.13) between $\mathcal{R}$ and $\mathcal{Q}$ is proportional to $\varphi(\bar{\mathbb{P}}; k)$. Our theorem also involves the quantity

$$b_{\max} := \max_{m=1,\dots,M} \left\| \int k_m(x,y) d\mathbb{P}_m(y) \right\|_\infty^2 .$$

Note that this is simply a constant whenever the kernels k_m are bounded.

Theorem 4.3.1 (Population control of subspaces)**.** *For any finite mixture $\bar{\mathbb{P}}$ with difficulty function bounded as $\varphi(\bar{\mathbb{P}}; k) \leq \left[576\sqrt{12 + b_{\max}}\right]^{-1} \Gamma^2_{\min}(\bar{\mathbb{P}})$, the distance between subspaces $\mathcal{Q}$ and $\mathcal{R}$ is bounded as*

$$\rho(\mathcal{Q}, \mathcal{R}) \leq 16\sqrt{12 + b_{\max}}\, \varphi(\bar{\mathbb{P}}; k). \tag{4.3.15}$$

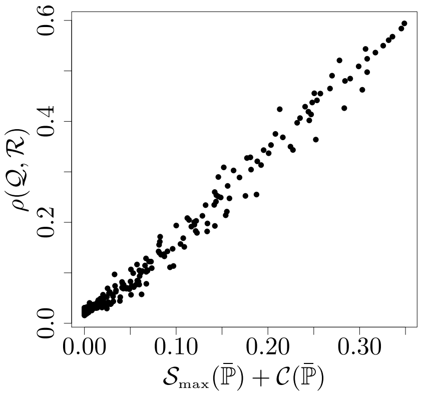

Figure 4.3. The ρ-distance between $\mathcal{Q}$ and $\mathcal{R}$ scales linearly with $\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})$. Each point corresponds to a different offset μ between the two Gaussian mixture components from Example 3.

The relationship (4.3.15) is easy to understand in the context of translated copies of identical mixture components. Consider the mixture of Gaussians with Gaussian kernel setup in Example 3. Recall from equation (4.3.10) that the indivisibility parameter is independent of the offset μ. Hence in this setting relationship (4.3.15) simplifies to

$$\rho(\mathcal{Q}, \mathcal{R}) \leq c[\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})]^{1/2}.$$

Figure 4.3 shows a clear linear relationship between $\rho(\mathcal{Q}, \mathcal{R})$ and $\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})$, suggesting that it might be possible to remove the square root in the clustering difficulty (4.3.14).

One important consequence of the relationship (4.3.15) stems from geometric structure in the square root kernelized density embedding. When there is little overlap between mixture components with respect to the kernel, the square root kernelized densities are not simultaneously large, i.e., $\Phi_{\mathcal{Q}}(X)$ will have at most one component much different from zero. Therefore the data will concentrate in tight spikes about the axes. This is illustrated in Figure 4.4.

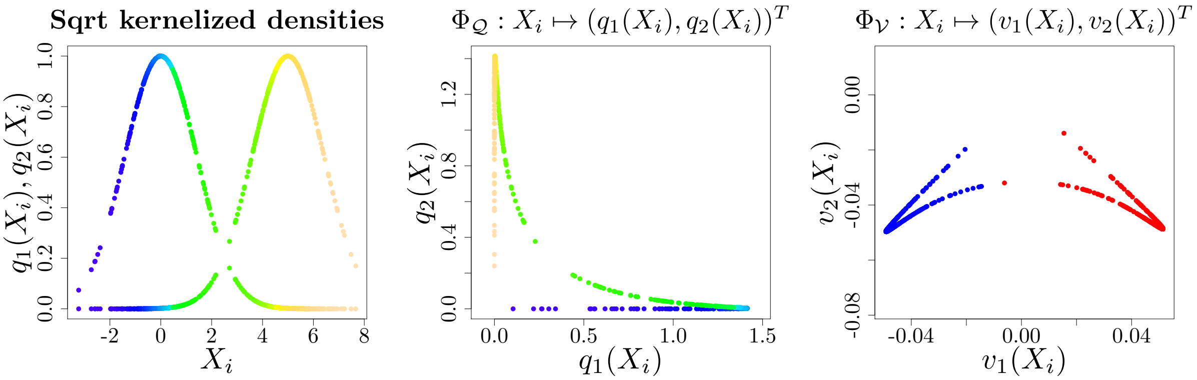

Figure 4.4. Geometric structure in the square root kernelized density embedding. (left) Square root kernelized densities for a mixture of Gaussians with Gaussian kernel. The color of the i-th dot indicates the likelihood ratio of the mixture components at X_i. (center) Data embedded under the square root kernelized density embedding $\Phi_{\mathcal{Q}}$, colored by likelihood ratio. (right) Normalized Laplacian embedding of the samples, colored by latent label.

Finite sample analysis

Thus far, our analysis has been limited to the population level, corresponding to the ideal case of infinitely many samples. We now turn to the case of finite samples. Here an additional level of analysis is required, in order to relate empirical versions (based on the finite collection of samples) to their population analogues. Doing so allows us to show that under suitable conditions, the Laplacian embedding applied to i.i.d. samples drawn from a finite mixture satisfies a certain geometric property, which we call *orthogonal cone structure*, or OCS for short.

We begin by providing a precise definition of when an embedding $\Phi : \mathcal{X} \to \mathbb{R}^M$ reveals orthogonal cone structure. Given a collection of labeled samples $\{X_i, Z_i\}_{i=1}^n \subset \mathcal{X} \times [M]$ drawn from a M-component mixture distribution, we let $\mathcal{Z}_m = \{i \in [n] \mid Z_i = m\}$ denote the subset of samples drawn from mixture component $m = 1, \ldots, M$. For any set $\mathcal{Z} \subseteq [n] = \{1, 2, \ldots, n\}$, we use $|\mathcal{Z}|$ to denote its cardinality. For vectors $u, v \in \mathbb{R}^n$, we use $\text{angle}(u, v) = \arccos \frac{\langle u, v \rangle}{\|u\|_2 \|v\|_2}$ to denote the angle between them. With this notation, we have the following:

Definition 3 (Orthogonal cone structure (OCS))**.** Given parameters $\alpha \in (0, 1)$ and $\theta \in (0, \frac{\pi}{4})$, the embedded data set $\{\Phi(X_i), Z_i\}_{i=1}^n$ has (α, θ)-OCS if there is an orthogonal basis $\{e_1, \ldots, e_M\}$ of $\mathbb{R}^M$ such that

$$\left| \left\{ i \in [n] \mid \text{angle}\big(\Phi(X_i), e_m\big) < \theta \right\} \cap \mathcal{Z}_m \right| \geq (1 - \alpha) \left| \mathcal{Z}_m \right| \qquad \text{for all } m = 1, \ldots, M.$$

In words, a labeled dataset has orthogonal cone structure if most pairs of embedded data points with distinct labels are almost orthogonal. See Figure 4.5 for an illustration of this property.

Our main theorem in the finite sample setting establishes that under suitable conditions, the normalized Laplacian embedding has orthogonal cone structure. In order to state this result precisely, we require a few additional conditions.

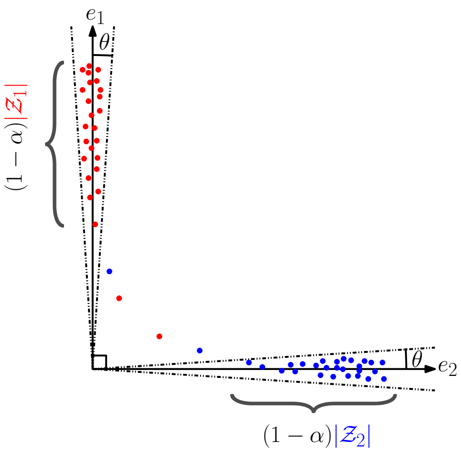

Figure 4.5. Visualizing (α, θ)-OCS: The labeled set of points plotted above has (α, θ) orthogonal cone structure with respect to its labeling. The color of each dot indicates the value of the corresponding label $Z_i \in \{1, 2\}$, where 1 corresponds to red and 2 to blue. This set of points has (α, θ)-orthogonal cone structure because a fraction $1 - \alpha$ of the red points (for which $Z_i = 1$) lie with an angle θ of e_1, a fraction $1 - \alpha$ blue points (for which $Z_i = 2$) lie with an angle θ of e_2, and e_1 is orthogonal to e_2.

Kernel parameters As a consequence of the compactness of $\mathcal{X}$, the kernel function is b-bounded, meaning that $k(x, x') \in (0, b)$ for all $x, x' \in \mathcal{X}$. As another consequence, the kernelized densities are lower bounded as $q_m(X^m) \geq r > 0$ with $\bar{\mathbb{P}}$-probability one. In the following statements, we use $c, c_0, c_1, \ldots$ to denote quantities that may depend on b, and r but are otherwise independent of the mixture distribution.

Tail decay The tail decay of the mixture components enters our finite sample result through the function $\psi : (0, \infty) \to [0, 1]$, defined by

$$\psi(t) := \sum_{m=1}^{M} \mathbb{P}_m \Big[\frac{q_m^2(X)}{\|q_m\|_{\bar{\mathbb{P}}}^2} < t \Big]. \tag{4.3.16}$$

Note that ψ is an increasing function with $\psi(0) = 0$. The rate of increase of ψ roughly measures the tail-decay of the square root kernelized densities. Intuitively, perturbations to the square root kernelized density embedding will have a greater effect on points closer to the origin.

Recall the population level clustering difficulty parameter $\varphi(\bar{\mathbb{P}}; k)$ previously defined in equation (4.3.14). Our theory requires that there is some $\delta > 0$ such that

$$\underbrace{\Big[\varphi(\bar{\mathbb{P}}; k) + \frac{1}{\Gamma_{\min}^2(\bar{\mathbb{P}})} \Big(\frac{1}{\sqrt{n}} + \delta \Big) \Big]}_{\varphi_n(\delta)} \leq c \, \Gamma_{\min}^2(\bar{\mathbb{P}}). \tag{4.3.17}$$

In essence, we assume that the indivisibility of the mixture components is not too small compared to the clustering difficulty.

With this notation, the following result applies to i.i.d. labeled samples $\{(X_i, Z_i)\}_{i=1}^n$ from a M-component mixture $\bar{\mathbb{P}}$.

Theorem 4.3.2 (Finite-sample angular structure)**.** *There are constants c, c_0, c_1, c_2 depending only on b and r such that for any $\delta \in (0, \frac{\|k\|_{\bar{\mathbb{P}}}}{b\sqrt{2\pi}})$ satisfying condition (4.3.17) and any $t > \frac{c_0}{w_{\min}^3}\sqrt{\varphi_n(\delta)}$, the embedded data set $\{\Phi_{\mathcal{V}}(X_i), Z_i)\}_{i=1}^n$ has (α, θ)-OCS with*

$$|\cos\theta| \le \frac{c_0\sqrt{\varphi_n(\delta)}}{w_{\min}^3 t - c_0\sqrt{\varphi_n(\delta)}} \quad \text{and} \quad \alpha \le \frac{c_1}{(w_{\min})^{\frac{3}{2}}}\varphi_n(\delta) + \psi(2t), \tag{4.3.18}$$

and this event holds probability at least $1 - 8M^2 \exp\left(\frac{-c_2\, n\delta^4}{\delta^2 + \mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})}\right)$.

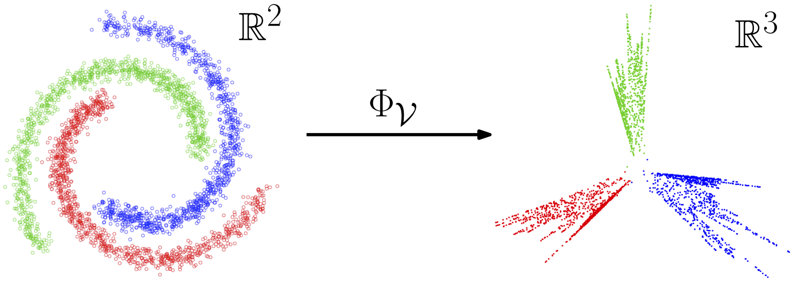

Figure 4.6. According to Theorem 4.3.2, the normalized Laplacian embedding of i.i.d. samples from a nonparametric mixture with small overlap, indivisible components, and large enough sample size, has (α, θ)-OCS with $\alpha \ll 1$ and $\theta \ll 1$. The left plot shows i.i.d. samples in $\mathbb{R}^2$, and the right plot displays the image (in $\mathbb{R}^3$) of these data under the normalized Laplacian embedding, $\Phi_{\mathcal{V}}$. The embedding was performed using a regularized Gaussian kernel. The color of each point indicates the latent label of that point.

Theorem 4.3.2 establishes that the embedding of i.i.d. samples from a finite mixture $\bar{\mathbb{P}}$ has orthogonal cone structure (OCS) *if* the components have small overlap and good indivisibility. This result holds with high probability on the sampling from $\bar{\mathbb{P}}$. See Figure 4.6 for an illustration of the theorem.

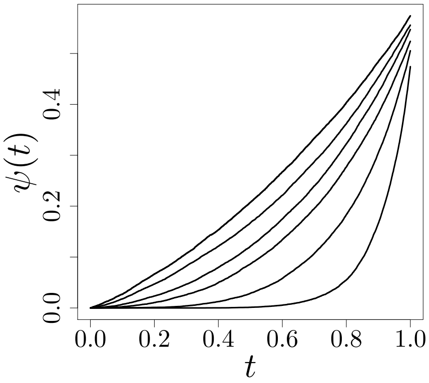

Figure 4.7. The tail decay function $\psi(t)$ roughly follows a power-law for the standard Gaussian distribution and Gaussian kernel with bandwidths $\nu \in \{0.15, 0.45, 0.75, 1, 1.5, 2.5\}$.

The tail decay of the mixture components enters the bounds on α and θ in different ways: the bound on θ is inversely proportional to t, but the bound on α is tighter for smaller t. Depending on how quickly ψ increases with t, it may very well be the dominant term in the bound on α. For example if there is a $\gamma > 0$ such that $\psi(t) \leq t^\gamma$ for all $t \in (0,1)$, and we set $t = \varphi_n^\beta$ for some $\beta \in (0,1)$, then we obtain the simplified bounds

$$|\cos\theta| \leq \frac{c}{w_{\min}^3} \varphi_n^{1/2-\beta} \quad \text{and} \quad \alpha \leq \frac{c}{(w_{\min})^{\frac{3}{2}}} \Big(\varphi_n + \varphi_n^{\gamma\beta} \Big).$$

Indeed, we find that whenever $\gamma\beta < 2$, the tail decay function is the dominant term in the bound on α. Note that this power-law tail decay is easy to verify for the Gaussian distribution with Gaussian kernel from Example 2 (see Figure 4.7).

Finally, the constants c, c_0, c_1, c_2 increase as the kernel bound b increases and as r decreases. This is where we need the tail truncation condition $r > 0$. This assumption is common in the literature (see Cao and Chen [31], for example). Both Luxburg, Belkin and Bousquet [77] and Rosasco, Belkin and De Vito [96] assume $k(x,y) \geq r > 0$, which is more restrictive. Note that this automatically holds if we add a positive constant to any kernel. This is sometimes called *regularization* and can significantly increase the performance of spectral clustering in practice [2].

Algorithmic consequences

In this section we apply our theory to study the performance of spectral clustering. The standard spectral clustering algorithm applies M-means to the embedded dataset. For completeness, we give pseudo code for the update step of M-means below.

```
$M$-means update
Input: Normalized embedded data $y_i := \frac{\Phi_\mathcal{V}(X_i)}{\|\Phi_\mathcal{V}(X_i)\|}$ for $i = 1, \dots, n$, and mean vectors $\{\mathbf{a}_1, \dots, \mathbf{a}_M\}$
for $m \in \{1, \dots, M\}$ do
    $\hat{\mathcal{Z}}_m \leftarrow \{ i : m = \operatorname{argmin}_\ell \|\mathbf{a}_\ell - y_i\| \}$
    $\mathbf{a}'_m \leftarrow \sum_{i \in \hat{\mathcal{Z}}_m} \frac{y_i}{|\hat{\mathcal{Z}}_m|}$
end for
return $\{\hat{\mathcal{Z}}_1, \dots, \hat{\mathcal{Z}}_M\}$ and $\{\mathbf{a}'_1, \dots, \mathbf{a}'_M\}$
```

In practice, we have found that applying M-means to an embedded dataset works well if the underlying orthogonal cone structure is "nice enough". The following proposition provides a quantitative characterization of this phenomenon. It applies to an embedded data set $\{\Phi_\mathcal{V}(X_i), Z_i)\}_{i=1}^n$ with (α,θ)-OCS, and an initialization of $\mathbf{a}_1, \dots, \mathbf{a}_M$ as uniformly random orthonormal vectors. Recall the notation $\mathcal{Z}_m = \{i \in [n] \mid Z_i = m\}$.

Proposition 4.3.3. *Suppose θ and α are sufficiently small that*

$$\frac{\alpha n + (1-\alpha)\left|\mathcal{Z}_m\right| \sin\theta}{(1-\alpha)\left|\mathcal{Z}_m\right|} \le \sin\frac{\pi}{8}, \quad \textit{and} \quad \frac{(1-\alpha)\left|\mathcal{Z}_m\right|\cos\theta - \alpha n}{\left|\mathcal{Z}_m\right| + \alpha n} \ge \frac{1}{2}, \tag{4.3.19}$$

for $m = 1, \ldots, M$. Then there is a constant c_K such that with probability at least $1 - \frac{4c_K\theta}{2\pi}$ over the random initialization, the M-means algorithm misclusters at most αn points. When $K = 2$, we have $c_K = 1$.

Intuitively, condition (4.3.19) requires α and θ to be small enough so that the different cones from the (α, θ)-OCS do not overlap.

Proof. We provide a detailed proof for the case $M = 2$. By the definition of (θ, α)-OCS, there exist orthogonal vectors e_1, e_2 such that a fraction $1-\alpha$ of the embedded samples with latent label m lie within an angle θ of e_m, $m = 1, 2$. Let us say that the initialization is *unfortunate* if some $\mathbf{a}_j$ falls within angle $\frac{\theta}{2}$ of the angular bisector of e_1 and e_2, an event which occurs with probability $\frac{4\theta}{2\pi}$.

Suppose without loss of generality that $\mathbf{a}_1$ is closer to e_1, and let $\mathbf{a}_1', \mathbf{a}_2'$ denote the updates

$$\mathbf{a}_m' = \sum_{i \in \hat{\mathcal{Z}}_m} \frac{v_i}{|\hat{\mathcal{Z}}_m|}, \qquad m = 1, 2.$$

If the initialization is *not* unfortunate, then all points in the θ-cone around e_1 are closer to $\mathbf{a}_1$ than $\mathbf{a}_2$. In this case, the (θ, α)-OCS implies that the e_2-coordinate of $\mathbf{a}_1'$ is at most

$$\frac{\alpha n + (1-\alpha)\left|\mathcal{Z}_1\right|, \sin\theta}{(1-\alpha)\left|\mathcal{Z}_1\right|} \le \sin\frac{\pi}{8},$$

and the e_1-coordinate of $\mathbf{a}_1'$ is at least

$$\frac{(1-\alpha)\left|\mathcal{Z}_1\right|\cos\theta - \alpha n}{\left|\mathcal{Z}_1\right| + \alpha n} \ge \frac{1}{2}.$$

We conclude that all points in the θ-cone about e_1 are closer to $\mathbf{a}_1'$ than $\mathbf{a}_2'$. Consequently, we find that after a single update step of M-means, all but a fraction α of the samples are correctly labeled. Moreover, this holds for all subsequent M-means updates. This completes the proof for $M = 2$.

The proof for general M follows the same steps. The probability that any $\mathbf{a}_m$ falls within angle $\frac{\theta}{2}$ of the angular bisector of any pair e_j, e_ℓ is still proportional to θ, with a constant of proportionality c_K that depends on M.

□

4.4 Proofs

We now turn to the proofs of our main results, beginning with the population level result stated in Theorem 4.3.1. We then provide the proof of Theorem 4.3.2 and Proposition 4.3.3.

Proof of Theorem 4.3.1

Our proof leverages an operator perturbation theorem due to Stewart [107] to show that $\mathcal{Q}$ is an approximate invariant subspace of the normalized Laplacian operator $\bar{\mathbf{T}}$ from equation (4.2.5). Recalling that $\Pi_{\mathcal{Q}}$ denotes the projection onto subspace $\mathcal{Q}$ (with $\Pi_{\mathcal{Q}^\perp}$ defined analogously), consider the following three operators

$$\mathbf{A} := \Pi_{\mathcal{Q}} \bar{\mathbf{T}}\, \Pi_{\mathcal{Q}}^*, \quad \mathbf{B} := \Pi_{\mathcal{Q}^\perp} \bar{\mathbf{T}}\, \Pi_{\mathcal{Q}^\perp}^*, \quad \text{and} \quad \mathbf{G} := \Pi_{\mathcal{Q}^\perp} \bar{\mathbf{T}}\, \Pi_{\mathcal{Q}}^*.$$

By definition, a subspace $\mathcal{Q}$ is invariant under $\bar{\mathbf{T}}$ if and only if $\mathbf{G} = 0$. In our setting, this ideal situation occurs when there is no overlap between mixture components. More generally, operator perturbation theory can be used guarantee that a space is approximately invariant as long as the Hilbert–Schmidt norm $|||\mathbf{G}|||_{\text{HS}}$ is not too large relative to the spectral separation between $\mathbf{A}$ and $\mathbf{B}$. In particular, define the quantities

$$\gamma := |||\mathbf{G}|||_{\text{HS}}, \quad \text{and} \quad \text{sep}(\mathbf{A}, \mathbf{B}) := \inf \big\{ |a - b| \mid a \in \sigma(\mathbf{A}), b \in \sigma(\mathbf{B}) \big\}.$$

In application to our problem, Theorem 3.6 of Stewart [107] guarantees that as long $\frac{\gamma}{\text{sep}(\mathbf{A},\mathbf{B})} < \frac{1}{2}$, then there is an operator $\mathbf{S} : \mathcal{Q} \to \mathcal{Q}^\perp$ such that

$$|||\mathbf{S}|||_{\text{HS}} \leq \frac{2\,\gamma}{\text{sep}(\mathbf{A}, \mathbf{B})} \tag{4.4.1}$$

such that $\text{Range}(\Pi_{\mathcal{Q}}^* + \Pi_{\mathcal{Q}^\perp}^*\, \mathbf{S})$ is an invariant subspace of $\bar{\mathbf{T}}$.

Accordingly, in order to apply this result, we first need to control the quantities $|||\mathbf{G}|||_{\text{HS}}$ and $\text{sep}(\mathbf{A}, \mathbf{B})$. The bulk of our technical effort is devoted to proving the following two lemmas:

Lemma 4.4.1 (Hilbert–Schmidt bound). *We have*

$$|||\mathbf{G}|||_{\text{HS}} \leq \frac{\sqrt{M(12 + b_{\max})}}{w_{\min}} \sqrt{\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})}, \tag{4.4.2}$$

where $b_{\max} := \max_{m=1,\ldots,M} \big\| \int k_m(x, y) d\mathbb{P}_m(y) \big\|_\infty^2$.

Lemma 4.4.2 (Spectral separation bound). *Under the hypothesis of the theorem, we have*

$$\sigma_{\min}(\mathbf{A}) \geq 1 - 13M[\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})]^{1/2}, \qquad \textit{and}$$
$$\sigma_{\max}(\mathbf{B}) \leq 1 - \frac{\Gamma^2}{8} + \frac{3[\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})]^{1/2}}{w_{\min}}.$$

Consequently, the spectral separation is lower bounded as

$$\text{sep}(\mathbf{A}, \mathbf{B}) \geq \frac{\Gamma^2}{16}. \tag{4.4.4}$$

See Appendices A.1 and A.2, respectively for the proof of these lemmas.

Combined with our earlier bound (4.4.1), these two lemmas guarantee that

$$\|\mathbf{S}\|_{\text{HS}} \leq 16\sqrt{12 + b_{\max}}\ \varphi(\bar{\mathbb{P}}, k). \tag{4.4.5}$$

Moreover, we find that $\text{Range}(\Pi_{\mathcal{Q}}^* + \Pi_{\mathcal{Q}^\perp}^*\ \mathbf{S})$ is equal to $\mathcal{R}$, the principal eigenspace of $\bar{\mathbf{T}}$. Indeed, by Stewart's theorem, the spectrum of $\bar{\mathbf{T}}$ is the disjoint union $\sigma(\bar{\mathbf{T}}) = \sigma(\mathbf{A} + \mathbf{G}^*\mathbf{S}) \cup \sigma(\mathbf{B} - \mathbf{S}\mathbf{G}^*)$. After some calculation using the upper bound on $\varphi(\bar{\mathbb{P}}, k)$ in the theorem hypothesis, we find that the spectrum of $\bar{\mathbf{T}}$ satisfies $\sigma_{\min}(\mathbf{A} + \mathbf{G}^*\mathbf{S}) > \sigma_{\max}(\mathbf{B} - \mathbf{S}\mathbf{G}^*)$, and any element $x \in \text{Range}(\Pi_{\mathcal{Q}}^* + \Pi_{\mathcal{Q}^\perp}^*\ \mathbf{S})$ must satisfy

$$\sup_{q \in \mathcal{Q}} \left\{ \frac{x^*\bar{\mathbf{T}}x}{x^*x} \,\middle|\, x = (\Pi_{\mathcal{Q}}^* + \Pi_{\mathcal{Q}^\perp}^*\ \mathbf{S})q \right\} > \sigma_{\max}(\mathbf{B} - \mathbf{S}\mathbf{G}^*).$$

Therefore $\text{Range}(\Pi_{\mathcal{Q}}^* + \Pi_{\mathcal{Q}^\perp}^*\ \mathbf{S}) = \mathcal{R}$.

The only remaining step is to translate the bound (4.4.5) into a bound on the norm $\|\Pi_{\mathcal{R}} - \Pi_{\mathcal{Q}}\|_{\text{HS}}$. Observe that the difference of projection operators can be written as

$$\Pi_{\mathcal{R}} - \Pi_{\mathcal{Q}} = (\Pi_{\mathcal{R}} + \Pi_{\mathcal{R}^\perp})(\Pi_{\mathcal{R}} - \Pi_{\mathcal{Q}}) = \Pi_{\mathcal{R}}\Pi_{\mathcal{Q}^\perp} - \Pi_{\mathcal{R}^\perp}\Pi_{\mathcal{Q}}.$$

Now Lemma 3.2 of Stewart [107] gives the explicit representations

$$\Pi_{\mathcal{R}} = (\mathbf{I} + \mathbf{S}^*\mathbf{S})^{-\frac{1}{2}}(\Pi_{\mathcal{Q}} + \mathbf{S}^*\Pi_{\mathcal{Q}^\perp}), \quad \text{and} \quad \Pi_{\mathcal{R}^\perp} = (\mathbf{I} + \mathbf{S}\mathbf{S}^*)^{-\frac{1}{2}}(\Pi_{\mathcal{Q}^\perp} + \Pi_{\mathcal{Q}}\mathbf{S}).$$

Consequently, we have

$$\|\Pi_{\mathcal{R}^\perp}\Pi_{\mathcal{Q}}\|_{\text{HS}} \leq \|(\mathbf{I} + \mathbf{S}\mathbf{S}^*)^{-\frac{1}{2}}\mathbf{S}\|_{\text{HS}}, \quad \text{and} \quad \|\Pi_{\mathcal{R}}\Pi_{\mathcal{Q}^\perp}\|_{\text{HS}} \leq \|(\mathbf{I} + \mathbf{S}^*\mathbf{S})^{-\frac{1}{2}}\mathbf{S}^*\|_{\text{HS}}.$$

By the continuous functional calculus (see §VII.1 of Reed and Simon [94]), we have the expansion

$$(\mathbf{I} + \mathbf{S}\mathbf{S}^*)^{-\frac{1}{2}} = \sum_{n=1}^{\infty} \binom{2n}{n} \frac{(\mathbf{S}\mathbf{S}^*)^{n-1}}{2^{2n}}.$$

Putting together the pieces, in terms of the shorthand $\epsilon = \|\mathbf{S}\|_{\text{HS}}$, we have

$$\|\Pi_{\mathcal{R}} - \Pi_{\mathcal{Q}}\|_{\text{HS}} \leq \frac{\epsilon}{2} \sum_{n=1}^{\infty} \binom{2n}{n} \left(\frac{\epsilon}{2}\right)^{2(n-1)} = \frac{2}{\epsilon}\left(\frac{1}{\sqrt{1-\epsilon^2}} - 1\right) \leq \epsilon,$$

which completes the proof.

Proof of Theorem 4.3.2

We say that a M-element subset (or M-tuple) of $\{X_1, \ldots, X_n\}$ is *diverse* if the latent labels of all points in the subset are distinct. Given some $\theta \in (0, \frac{\pi}{4})$, a M-tuple is θ-orthogonal if all its distinct pairs, when embedded, are orthogonal up to angle $\frac{\theta}{2}$. In order to establish (α, θ)-angular structure in the normalized Laplacian embedding of $\{X_1, \ldots, X_n\}$, we must show that there is a subset of $\{X_1, \ldots, X_n\}$ with at least $(1 - \alpha)n$ elements, and with the property that every diverse M-tuple from the subset is θ-orthogonal.

We break the proof into two steps. We first lower bound the total number of M-tuples that are diverse and θ-orthogonal. In the second step we construct the desired subset. We present the first step below and defer the second step to Appendix B.1 in the supplement.

Step 1: Consider a diverse M-tuple $(X^1, \ldots, X^M)$ constructed randomly by selecting X^m uniformly at random from the set $\{X_i \mid Z_i = m\}$ for $m = 1, \ldots, M$. Form the $M \times M$ random matrix

$$V = \begin{bmatrix} | & & | \\ \Phi_{\mathcal{V}}(X^1) & \cdots & \Phi_{\mathcal{V}}(X^M) \\ | & & | \end{bmatrix},$$

where $\Phi_{\mathcal{V}}$ denotes the normalized Laplacian embedding from equation (4.2.3). Let $\tilde{V}$ denote an independent copy of V. Let $Q \in \mathbb{R}^{M\times M}$ denote the diagonal matrix with entries $Q_{mm} = \frac{q_m(X^m)}{\|q_m\|_{\bar{\mathbb{P}}_n}}$, where $\bar{\mathbb{P}}_n$ is the empirical distribution over the samples $X_1, \ldots, X_n$, and define $Q_{\max} := \max_m \frac{\|q_m\|_\infty}{\|q_m\|_{\bar{\mathbb{P}}}}$.

At the core of our proof lies the following claim involving a constant c_3. For at least a fraction $1 - \frac{2Mc_3\varphi_n(\delta)}{\sqrt{w_{\min}}}$ of the diverse M-tuples, we have the inequality

$$\|V^T\tilde{V} - Q^2\|_{\text{HS}} \le \frac{32\sqrt{3}}{w_{\min}^3} Q_{\max}^2 \sqrt{c_3}\sqrt{\varphi_n(\delta)}, \tag{4.4.6}$$

holding on a high probability set $\mathcal{A}$. For the moment, we take this claim as given, before returning to define $\mathcal{A}$ explicitly and prove the claim.

When the inequality (4.4.6) is satisfied, we obtain the following upper bound on the off-diagonal elements of $V^T\tilde{V}$:

$$\left(V^T\tilde{V}\right)_{m\ell} \le \frac{32\sqrt{3}Q_{\max}^2}{w_{\min}^3}\sqrt{c_3}\sqrt{\varphi_n(\delta)} \quad \text{for } m \ne \ell.$$

This is useful because

$$\cos\text{angle}(\Phi_{\mathcal{V}}(X^m), \Phi_{\mathcal{V}}(X^\ell)) = \frac{(V^T\tilde{V})_{m\ell}}{\sqrt{(V^TV)_{mm}(V^T\tilde{V})_{\ell\ell}}}.$$

However, we must also lower bound $\min_m (V^T\tilde{V})_{mm}$. To this end by union bound we obtain

$$\Pr\left\{\min_m Q_{mm}^2 \le t\right\} = \Pr\left\{\min_m \frac{q_m^2(X^m)}{\|q_m\|_{\bar{\mathbb{P}}_n}^2} \le t\right\} \le \sum_{m=1}^M \Pr\left\{\frac{q_m^2(X^m)}{\|q_m\|_{\bar{\mathbb{P}}_n}^2} \le t\right\} := \psi_n(t). \tag{4.4.7}$$

On the set $\mathcal{A}_\psi := \{\sup_t |\psi_n(t) - \psi(t)| \le \delta\} \subset \mathcal{A}$, we may combine equations (4.4.6) and (4.4.7) to obtain

$$\min_m \left(V^T\tilde{V}\right)_{mm} \ge t - \frac{32\sqrt{3}Q_{\max}^2}{w_{\min}^3}\sqrt{c_3}\sqrt{\varphi_n(\delta)},$$

with probability at least $1 - \psi(2t)$. Therefore, there is a θ satisfying

$$|\cos\theta| \le \frac{32\sqrt{3}Q_{\max}^2\sqrt{c_3}\sqrt{\varphi_n(\delta)}}{w_{\min}^3 t - 32\sqrt{3}Q_{\max}^2\sqrt{c_3}\sqrt{\varphi_n(\delta)}}$$

such that at least a fraction $1 - \frac{2Mc_3\varphi_n(\delta)}{\sqrt{w_{\min}}} - \psi(2t)$ of the diverse M-tuples are θ-orthogonal on the set $\mathcal{A}$. This establishes the finite sample bound (4.3.18) with $c_0 := 2c_3$ and $c_1 := 32\sqrt{3}Q_{\max}^2\sqrt{c_3}$.

It remains to prove the intermediate claim (4.4.6). Define the matrix

$$A := \begin{bmatrix} \left\langle \frac{q_1}{\|q_1\|_{\bar{\mathbb{P}}_n}}, v_1 \right\rangle_{\bar{\mathbb{P}}_n} & \cdots & \left\langle \frac{q_1}{\|q_1\|_{\bar{\mathbb{P}}_n}}, v_M \right\rangle_{\bar{\mathbb{P}}_n} \\ \vdots & \ddots & \vdots \\ \left\langle \frac{q_M}{\|q_M\|_{\bar{\mathbb{P}}_n}}, v_1 \right\rangle_{\bar{\mathbb{P}}_n} & \cdots & \left\langle \frac{q_M}{\|q_M\|_{\bar{\mathbb{P}}_n}}, v_M \right\rangle_{\bar{\mathbb{P}}_n} \end{bmatrix}. \tag{4.4.8}$$

Note that the entries of AA^T are

$$\left(AA^T\right)_{m\ell} = \frac{\langle \Pi_{\mathcal{V}} q_m, \Pi_{\mathcal{V}} q_\ell \rangle_{\bar{\mathbb{P}}_n}}{\|q_m\|_{\bar{\mathbb{P}}_n} \|q_\ell\|_{\bar{\mathbb{P}}_n}}.$$

The off-diagonal elements satisfy

$$\left(AA^T\right)_{m\ell} \le 3(\hat{\varphi} + \sqrt{\hat{\mathcal{S}}_{\max}}), \quad \text{for } m \neq \ell$$

where $\hat{\varphi} = \max_m \frac{\|q_m - \Pi_{\mathcal{V}} q_m\|_{\bar{\mathbb{P}}_n}}{\|q_m\|_{\bar{\mathbb{P}}_n}}$, and $\hat{\mathcal{S}}_{\max} = \max_{m\neq\ell} \frac{\|q_\ell\|_{\mathbb{P}_m^n}^2}{\|q_m\|_{\bar{\mathbb{P}}_n}^2}$ (and $\mathbb{P}_m^n$ denotes the empirical distribution for the samples with latent label m). Similarly, the diagonal elements satisfy $\left|\left(AA^T\right)_{mm} - 1\right| \le 3\hat{\varphi}$. Putting together the pieces yields $\|AA^T - I\|_{\text{HS}}^2 \le 3M^2(\hat{\varphi} + \hat{\mathcal{S}}_{\max})$, which in turn implies

$$\|\left(AA^T\right)^{-1} - I\|_{\text{HS}}^2 \le \frac{3M^2(\hat{\varphi} + \sqrt{\hat{\mathcal{S}}_{\max}})}{1 - 3M^2(\hat{\varphi} + \sqrt{\hat{\mathcal{S}}_{\max}})}.$$

We now transform this inequality into one involving $V^T\tilde{V}$. Write $B = AV$, and $\tilde{B} = A\tilde{V}$ note that $V^T\tilde{V} = B^T\left(AA^T\right)^{-1}\tilde{B}$. Therefore, we find that

$$\begin{aligned} \|V^T\tilde{V} - Q^2\|_{\text{HS}} &\le \|B^T\tilde{B} - Q^2\|_{\text{HS}} + \|B^T\left[\left(AA^T\right)^{-1} - I\right]\tilde{B}\|_{\text{HS}} \\ &\le 3\|Q\|_{\text{HS}}\|B - Q\|_{\text{HS}} + \|B\|_{\text{HS}}^2\|\left(AA^T\right)^{-1} - I\|_{\text{HS}}, \end{aligned}$$

where the last inequality used $\|B\|_{\text{HS}} \le 2\|Q\|_{\text{HS}}$. Now note that the entries of B are $B_{m\ell} = \frac{\Pi_{\mathcal{V}} q_m(X^\ell)}{\|q_m\|_{\bar{\mathbb{P}}_n}}$. Therefore the difference $B - Q$ satisfies

$$\mathbb{E}\left[\|B - Q\|_{\text{HS}}^2 | X_1, \ldots, X_n\right] \le M^2\left(\frac{\hat{\varphi}}{\sqrt{\hat{w}_{\min}}} + \sqrt{\hat{\mathcal{S}}_{\max}}\right)^2 + M\frac{\hat{\varphi}^2}{\hat{w}_{\min}}, \tag{4.4.9}$$

where $\hat{w}_{\min} = \min_m \frac{n_m}{n}$, and the expectation above is over the selection of the random M-tuple $(X^1, \ldots, X^M)$.[4]

[4] Note that are two different types of randomness at play in the construction of V, and hence B; there is randomness in the generation of the i.i.d. samples $X_1, \ldots, X_n$ from $\bar{\mathbb{P}}$, and there is randomness in the selection of the diverse M-tuple $(X^1, \ldots, X^M)$.

Both $\hat{\mathcal{S}}_{\max}$ and $\hat{\varphi}$ are small with high probability. Indeed, Bernstein's inequality guarantees that

$$\sqrt{\hat{\mathcal{S}}_{\max}} \leq \sqrt{\mathcal{S}_{\max}} + \delta \tag{4.4.10}$$

with probability at least $1 - 2M^2 \exp \frac{-n(\mathcal{S}_{\max}+\delta^2)^2}{8Q_{\max}^2(2\mathcal{S}_{\max}+\delta^2)}$. We control $\hat{\varphi}$ with a finite sample version of Theorem 4.3.1, which we state as Proposition 4.4.3 below.

Let $\mathcal{V} = \text{span}\{v_1, \ldots, v_M\}$ denote the principal eigenspace of the normalized Laplacian matrix.

Proposition 4.4.3. *There are constants c_2', c_3 such that for any $\delta \in (0, \frac{\|k\|_{\bar{\mathbb{P}}}}{b\sqrt{2\pi}})$ satisfying condition (4.3.17), we have*

$$\hat{\varphi} \leq c_3 \varphi_n(\delta) \tag{4.4.11}$$

with probability at least $1 - 10M \exp\big(\frac{-nc_2'\delta^4}{\delta^2+\mathcal{S}_{\max}(\bar{\mathbb{P}})+\mathcal{C}(\bar{\mathbb{P}})}\big)$.

See Section 4.4 for the proof of this auxiliary result.

On the set $\{\hat{\varphi} \leq c_3\varphi_n(\delta)\} \cap \{\hat{w}_{\min} \geq \frac{1}{2} w_{\min}\} := \mathcal{A}_\zeta \cap \mathcal{A}_w$, Equation (4.4.9) simplifies to

$$\mathbb{E}\big[\|B - Q\|_{\text{HS}}^2 | X_1, \ldots, X_n\big] \leq \frac{4M^2 c_3^2 \varphi_n^2(\delta)}{w_{\min}},$$

whenever $\big(\frac{\sqrt{2}c_3\varphi_n(\delta)}{\sqrt{w_{\min}}} + \mathcal{S}_{\max} + \delta\big)^2 \leq \frac{3c_3^2\varphi_n^2(\delta)}{w_{\min}}$, which is a consequence of condition (4.3.17). By Markov's inequality we obtain the following result: at least a fraction $1 - \frac{2Mc_3\varphi_n(\delta)}{\sqrt{w_{\min}}}$ of the diverse M-tuples satisfy

$$\|B - Q\|_{\text{HS}}^2 \leq \frac{2Mc_3\varphi_n}{\sqrt{w_{\min}}}. \tag{4.4.12}$$

For the diverse M-tuples that do satisfy inequality (4.4.12) we find that

$$\|V^T\tilde{V} - Q^2\|_{\text{HS}} \leq \big(\frac{6\sqrt{2}M^{3/2}Q_{\max}}{w_{\min}^{1/4}} + 32\sqrt{3}M^3Q_{\max}^2\big)\sqrt{c_3}\sqrt{\varphi_n},$$

valid on the set $\mathcal{A} = \mathcal{A}_w \cap \mathcal{A}_q \cap \mathcal{A}_\psi \cap \{\hat{\varphi} \leq c_3\varphi_n(\delta)\} \cap \{\sqrt{\hat{\mathcal{S}}_{\max}} \leq \sqrt{\mathcal{S}_{\max}} + \delta\}$, thereby establishing the bound (4.4.6).

To complete the first step of the proof of Theorem 4.3.2, it remains to control the probability of $\mathcal{A}$. By Hoeffding's inequality, we have $\mathbb{P}[\mathcal{A}_w] \geq 1 - Me^{\frac{-nw_{\min}^2}{2}}$. Finally, an application of Bernstein's inequality controls the probability of $\mathcal{A}_q$, and an application of Glivenko Cantelli controls the probability of $\mathcal{A}_\psi$. Putting together the pieces we find that $\mathcal{A}$ holds with probability at least $1 - 8M^2 \exp\big(\frac{-nc_2\delta^4}{\delta^2+\mathcal{S}_{\max}(\bar{\mathbb{P}})+\mathcal{C}(\bar{\mathbb{P}})}\big)$, where $c_2 := \min\big(c_2', \frac{1}{8Q_{\max}^2}\big)$.

Proof of Proposition 4.4.3

Consider the operator $\hat{\mathbf{T}} : L^2(\bar{\mathbb{P}}_n) \to L^2(\bar{\mathbb{P}}_n)$ defined by

$$(\hat{\mathbf{T}} f)(x) = \int \frac{1}{\bar{q}_n(x)} k(x,y) \frac{f(y)}{\bar{q}_n(y)} d\bar{\mathbb{P}}_n(y),$$

where $\bar{q}_n(x) = \frac{1}{n}\sum_{i=1}^{M} k(X_i, x)$ is the square root kernelized density for the empirical distribution $\bar{\mathbb{P}}_n$ over the data $X_1, \ldots, X_n$. $\bar{k}^n(x,y) := \frac{k(x,y)}{\bar{q}_n(x)\bar{q}_n(y)}$ for the normalized kernel function. Note that for any $f \in L^2(\bar{\mathbb{P}}_n)$ and $v \in \mathbb{R}^n$ with coordinates $v_i = f(X_i)$, we have $(\hat{\mathbf{T}} f)(X_j) = (Lv)_j$, where L is the normalized Laplacian matrix (4.2.2). Consequently, the principal eigenspace $\mathcal{V}$ of L is isomorphic to the principal eigenspace of $\hat{\mathbf{T}}$ which we also denote by $\mathcal{V}$ for simplicity.

To prove the proposition, we must relate $\hat{\mathbf{T}}$ to the normalized Laplacian operator $\bar{\mathbf{T}}$. These operators differ in both their measures of integration—namely, $\bar{\mathbb{P}}_n$ versus $\bar{\mathbb{P}}$—and their kernels, namely $\frac{k(x,y)}{\bar{q}_n(x)\bar{q}_n(y)}$ versus $\frac{k(x,y)}{\bar{q}(x)\bar{q}(y)}$. To bridge the gap we introduce an intermediate operator $\tilde{\mathbf{T}} : L^2(\bar{\mathbb{P}}_n) \to L^2(\bar{\mathbb{P}}_n)$ defined by

$$(\tilde{\mathbf{T}} f)(x) = \int \frac{1}{\bar{q}(x)} k(x,y) \frac{f(y)}{\bar{q}(y)} d\bar{\mathbb{P}}_n(y).$$

Let $\tilde{\mathcal{V}}$ denote the principal eigenspace of $\tilde{\mathbf{T}}$. The following lemma bounds the ρ-distance between the principal eigenspaces of $\tilde{\mathbf{T}}$ and $\hat{\mathbf{T}}$.

Lemma 4.4.4. *For any $\delta \in [0, \frac{\|k\|_{\bar{\mathbb{P}}}}{b\sqrt{2\pi}}]$ satisfying condition (4.3.17), we have*

$$\rho(\mathcal{V}, \tilde{\mathcal{V}}) \leq \frac{c_4}{\Gamma^2}\Big(\frac{1}{\sqrt{n}} + \delta\Big), \tag{4.4.13}$$

with probability at least $1 - 6e^{\frac{-n\pi\delta^2}{2}}$, where $c_4 = 1024\sqrt{2\pi}\frac{\|k\|_{\bar{\mathbb{P}}} b}{r^4}$.

See Appendix B.3 for a proof of this lemma.

We must upper bound $\|q_m - \Pi_{\mathcal{V}} q_m\|_{\bar{\mathbb{P}}_n}$. By triangle inequality,

$$\begin{aligned} \|q_m - \Pi_{\mathcal{V}} q_m\|_{\bar{\mathbb{P}}_n} \leq \|q_m - \Pi_{\mathcal{R}} q_m\|_{\bar{\mathbb{P}}_n} &+ \|\Pi_{\mathcal{R}} q_m - \Pi_{\tilde{\mathcal{V}}} q_m\|_{\bar{\mathbb{P}}_n} \\ &+ \|\Pi_{\mathcal{V}} q_m - \Pi_{\tilde{\mathcal{V}}} q_m\|_{\bar{\mathbb{P}}_n}. \end{aligned}$$

Note that $\|\Pi_{\mathcal{V}} q_m - \Pi_{\tilde{\mathcal{V}}} q_m\|_{\bar{\mathbb{P}}_n} \leq \|q_m\|_{\bar{\mathbb{P}}_n} \rho(\tilde{\mathcal{V}}, \mathcal{V})$. We can control this term with the lemma. The term $\|q_m - \Pi_{\mathcal{R}} q_m\|_{\bar{\mathbb{P}}_n}$ is the empirical version of a quantity controlled by Theorem 4.3.1. We handle the empirical fluctuations with a version of Bernstein's inequality. For $\delta_p \geq 0$ we have the inequality

$$\|q_m - \Pi_{\mathcal{R}} q_m\|_{\bar{\mathbb{P}}_n} \leq \|q_m - \Pi_{\mathcal{R}} q_m\|_{\bar{\mathbb{P}}} + \delta_p \tag{4.4.14}$$

with probability at least $1 - 2\exp\big(-\frac{n\delta_p^4}{8(\delta_p^2 + c_{\text{pop}}^2\varphi^2)\tilde{Q}_{\max}^2}\big)$, where $\tilde{Q}_{\max} = \max_m \frac{\|q_m - \Pi_{\mathcal{R}} q_m\|_\infty}{\|q_m\|_{\bar{\mathbb{P}}}}$ and $c_{\text{pop}} := 16\sqrt{12 + \frac{b_{\max}}{M}}$.

It remains to control $\|\Pi_{\mathcal{R}} q_m - \Pi_{\tilde{\mathcal{V}}} q_m\|_{\bar{\mathbb{P}}_n}$. Let $\bar{\mathcal{H}}$ denote the reproducing kernel Hilbert space (RKHS)[5] for the kernel $\bar{k}$. Now we define two integral operators on $\bar{\mathcal{H}}$. Let $\bar{\mathbf{H}}$ denote the operator defined by

$$(\bar{\mathbf{H}}h)(x) = \int \bar{k}(x,y)h(y)d\bar{\mathbb{P}}(y),$$

and similarly let $\tilde{\mathbf{H}} : \bar{\mathcal{H}} \to \bar{\mathcal{H}}$ denote the operator defined by

$$(\tilde{\mathbf{H}}h)(x) = \int \bar{k}(x,y)h(y)d\bar{\mathbb{P}}_n(y).$$

Both $\bar{\mathbf{H}}$ and $\tilde{\mathbf{H}}$ are self-adjoint, compact operators on $\bar{\mathcal{H}}$ and have real, discrete spectra. Let $\mathcal{G}$ denote the principal M-dimensional eigenspace of $\bar{\mathbf{H}}$ and let $\tilde{\mathcal{G}}$ denote the principal M-dimensional principal eigenspace of $\tilde{\mathbf{H}}$. The following lemma bounds the ρ-distance between these subspaces of $\bar{\mathcal{H}}$.

Lemma 4.4.5. *For any $\delta > 0$ satisfying condition (4.3.17), we have*

$$\rho(\mathcal{G}, \tilde{\mathcal{G}}) \le \frac{c_5}{\Gamma^2}\Big(\frac{1}{\sqrt{n}} + \delta\Big)$$

with probability at least $1 - 2e^{-n\pi\mathbb{E}\bar{k}(\bar{X},\bar{X})\delta^2}$, where $c_5 = 64\sqrt{2\pi}\sqrt{\mathbb{E}\bar{k}(\bar{X},\bar{X})}\frac{b}{r^2}$.

See Appendix B.2 for the proof of this Lemma.

By the triangle inequality, we have

$$\|\Pi_{\mathcal{R}} q_m - \Pi_{\tilde{\mathcal{V}}} q_m\|_{\bar{\mathbb{P}}_n} \le \|\Pi_{\mathcal{R}} q_m - \Pi_{\mathcal{G}} q_m\|_{\bar{\mathbb{P}}_n} + \left\|\Pi_{\mathcal{G}} q_m - \Pi_{\tilde{\mathcal{G}}} q_m\right\|_{\bar{\mathbb{P}}_n} + \left\|\Pi_{\tilde{\mathcal{G}}} q_m - \Pi_{\tilde{\mathcal{V}}} q_m\right\|_{\bar{\mathbb{P}}_n}.$$

We claim that

$$\|\Pi_{\mathcal{R}} q_m - \Pi_{\mathcal{G}} q_m\|_{\bar{\mathbb{P}}_n} = 0, \quad \text{and} \quad \left\|\Pi_{\tilde{\mathcal{G}}} q_m - \Pi_{\tilde{\mathcal{V}}} q_m\right\|_{\bar{\mathbb{P}}_n} = 0. \tag{4.4.15}$$

We take these identities as given for the moment, before returning to prove them at the end of this subsection.

Now the term $\left\|\Pi_{\mathcal{G}} q_m - \Pi_{\tilde{\mathcal{G}}} q_m\right\|_{\bar{\mathbb{P}}_n}$ can be controlled using the lemma in the following way. For any $h \in \bar{\mathcal{H}}$, note that

$$\|h\|^2_{\bar{\mathbb{P}}_n} = \frac{1}{n}\sum_{i=1}^{n} \left\langle h, \bar{k}_{X_i}\right\rangle^2_{\bar{\mathcal{H}}} \le \frac{1}{n}\sum_{i=1}^{n} \|h\|^2_{\bar{\mathcal{H}}}\, \bar{k}(X_i, X_i)$$

by Cauchy-Schwarz for the RKHS inner product. Using this logic with $h = \Pi_{\mathcal{G}} q_m - \Pi_{\tilde{\mathcal{G}}} q_m$, we find

$$\left\|\Pi_{\mathcal{G}} q_m - \Pi_{\tilde{\mathcal{G}}} q_m\right\|_{\bar{\mathbb{P}}_n} \le \|q_m\|_{\bar{\mathcal{H}}} \sqrt{\frac{1}{n}\sum_{i=1}^{n} \bar{k}(X_i, X_i)}\rho(\mathcal{G}, \tilde{\mathcal{G}}). \tag{4.4.16}$$

[5] We give a brief introduction to the theory of reproducing kernel Hilbert spaces and provide some references for further reading on the subject in Appendix C.2.

Collecting our results and applying Lemmas 4.4.4 and 4.4.5 yields

$$\|q_m - \Pi_{\mathcal{V}} q_m\|_{\bar{\mathbb{P}}_n} \leq \big(c_{\text{pop}}\varphi + \delta_p\big)\|q_m\|_{\bar{\mathbb{P}}} + \frac{c_n \|q_m\|_{\bar{\mathbb{P}}_n}}{\Gamma^2}\big(\frac{1}{\sqrt{n}} + \delta\big),$$

where

$$c_n := \frac{256\sqrt{2\pi}b}{r^2}\big[\frac{\|q_m\|_{\bar{\mathcal{H}}}}{\|q_m\|_{\bar{\mathbb{P}}}}\mathbb{E}\bar{k}(\bar{X},\bar{X}) + \frac{2}{r^2}\big]. \tag{4.4.17}$$

By an application of Bernstein's inequality, we have

$$\|q_m\|_{\bar{\mathbb{P}}_n} \leq \sqrt{2}\,\|q_m\|_{\bar{\mathbb{P}}}$$

with probability at least $1 - 2e^{\frac{-n}{16Q_{\max}^2}}$. For $\delta \in (0, \frac{1}{2\sqrt{2\pi}Q_{\max}})$, we have

$$2e^{\frac{-n{c_{\text{pop}}}^2\delta^4}{8\Gamma^4\tilde{Q}_{\max}^2(\delta^2+\mathcal{S}_{\max}(\bar{\mathbb{P}})+\mathcal{C}(\bar{\mathbb{P}}))}} + 6e^{\frac{-n\pi\delta^2}{2}} + 2e^{\frac{-n}{16Q_{\max}^2}} \leq 10e^{-\frac{nc_2'\delta^4}{\delta^2+\mathcal{S}_{\max}(\bar{\mathbb{P}})+\mathcal{C}(\bar{\mathbb{P}})}},$$

where $\delta_p = \frac{c_{\text{pop}}\delta}{\Gamma^2}$, and $c_2' = \min\big(\frac{{c_{\text{pop}}}^2}{8\Gamma^4\tilde{Q}_{\max}^2}, \frac{\pi}{2}\big)$. Modulo the claim, this proves the proposition with $c_3 = 2\max(c_{\text{pop}}, c_n)$.

We now return to prove the claim (4.4.15). Note the following relation between the eigenfunctions of $\bar{\mathbf{T}}$ and those of $\bar{\mathbf{H}}$: if r_i is an eigenfunction of $\bar{\mathbf{T}}$ with eigenvalue λ_i and $\|r_i\|_{\bar{\mathbb{P}}} = 1$, then $g_i := \sqrt{\lambda_i}r_i$ has unit norm in $\bar{\mathcal{H}}$, and is an eigenfunction of $\bar{\mathbf{H}}$ with eigenvalue λ_i. Note that the eigenfunctions r_i of $\bar{\mathbf{T}}$ form an orthonormal basis of $L^2(\bar{\mathbb{P}})$, and therefore $q_m = \sum_{i=1}^{\infty} a_i r_i$, where a_i are the coefficients $\langle q_m, r_i\rangle_{\bar{\mathbb{P}}}$. By the observation above, we have the equivalent representation $q_m = \sum_{i=1}^{\infty} \frac{a_i}{\sqrt{\lambda_i}} g_i$. Therefore the $L^2(\bar{\mathbb{P}})$ projection onto $\mathcal{R} = \text{span}\{r_1, \ldots, r_M\}$ is $\Pi_{\mathcal{R}} q_m = \sum_{i=1}^{M} a_i r_i$, and the $\bar{\mathcal{H}}$ projection onto $\mathcal{G} = \text{span}\{g_1, \ldots, g_M\}$ is $\Pi_{\mathcal{G}} q_m = \sum_{i=1}^{M} \frac{a_i}{\sqrt{\lambda_i}} g_i$. Therefore the relation $g_i = \sqrt{\lambda_i}r_i$ implies $\|\Pi_{\mathcal{R}} - \Pi_{\mathcal{G}}\|_{\bar{\mathbb{P}}_n} = 0$. Similar reasoning yields $\big\|\Pi_{\tilde{\mathcal{G}}} q_\ell - \Pi_{\tilde{\mathcal{V}}} q_\ell\big\|_{\bar{\mathbb{P}}_n} = 0$.

4.5 Discussion

In this paper, we have analyzed the performance of spectral clustering in the context of nonparametric finite mixture models. Our first main contribution is an upper bound on the distance between the population level normalized Laplacian embedding and the square root kernelized density embedding. This bound depends on the maximal similarity index, the coupling parameter, and the indivisibility parameter. These parameters all depend on the kernel function, and we present our analysis for a fixed but arbitrary kernel.

Although this dependence on the kernel function might seem undesirable, it is actually necessary to guarantee identifiability of the mixture components in the following sense. A mixture with fully nonparametric components is a very rich model class: without any restrictions on the mixture components, any distribution can be written as a M-component mixture in uncountably many ways. Conversely, when the clustering difficulty function is zero, the representation of a distribution as a mixture is unique. In principle, one could optimize over the convex cone of symmetric positive definite kernel functions so to minimize our clustering difficulty parameter. In our preliminary numerical experiments, we have found promising results in using this strategy to choose the bandwidth in a family of kernels.

Building on our population-level result, we also provided a result that characterizes the normalized Laplacian embedding when applied to a finite collection of n i.i.d. samples. We find that when the clustering difficulty is small, the embedded samples take on approximate orthogonal structure: samples from different components are almost orthogonal with high probability. The emergence of this form of angular structure allows an angular version of M-means to correctly label most of the samples.

Perhaps surprising is the fact that the optimal bandwidth (minimizing our upper bound) is non-zero. Although we only provide an upper bound, we believe this is fundamental to spectral clustering, not an artifact of our analysis. Again, the principal M-dimensional eigenspace of the Laplacian operator is not a well-defined mathematical object when the bandwidth is zero. Indeed, as the bandwidth shrinks to zero, the eigengap distinguishing this eigenspace from the remaining eigenfucntion vanishes. This eigenspace, however, is the population-level version of the subspace onto which spectral clustering projects. For this reason, we caution against shrinking the bandwidth indefinitely to zero, and we conjecture that there is an optimal population level bandwidth for spectral clustering. However, we should mention that we cannot provably rule out the optimality of an appropriately slowly shrinking bandwidth, and we leave this to future work. Further investigation of kernel bandwidth selection for spectral clustering is an interesting avenue for future work.

4.6 Supplementary proofs for Theorem 1

This appendix is devoted to the proofs of Lemmas 1 and 2. In addition to the normalized Laplacian operator $\bar{\mathbf{T}}$, our proofs involve several other quantities, which we introduce here. For a general distribution $\mathbb{P}$ and kernel k, we define the kernel integral operator

$$\mathbf{T}_{k,\mathbb{P}} : f \mapsto \int f(x)k(x,\cdot)d\mathbb{P}(x).$$

For $m = 1, \ldots, M$, we define the difference operator

$$E_m f = \mathbf{T}_{k_m,\mathbb{P}_m} f - w_m \mathbf{T}_{\bar{k},\mathbb{P}_m} f. \tag{4.6.1}$$

Note that by definition, the operator E_m an integral operator with kernel $k_m - w_m\bar{k}$, and hence we have the bound

$$|||E_m|||_{\text{op}} \leq \left\|k_m - w_m\bar{k}\right\|_{\mathbb{P}_m} \qquad \text{for } m = 1, \ldots, M.$$

This bound is useful in the analysis to follow.

Proof of Lemma 1

We denote the normalized version of q_m by $\tilde{q}_m = \frac{q_m}{\|q_m\|_{\bar{\mathbb{P}}}}$. By Jensen's inequality and convexity of the squared HS-norm,

$$|||\Pi_{\mathcal{Q}^\perp}\bar{\mathbf{T}}\Pi_{\mathcal{Q}}^*|||_{\text{HS}}^2 \leq \sum_{m=1}^{M} w_m |||\Pi_{\mathcal{Q}^\perp}\mathbf{T}_{\bar{k},\mathbb{P}_m}\Pi_{\mathcal{Q}}^*|||_{\text{HS}}^2 \leq M \max_m w_m |||\Pi_{\mathcal{Q}^\perp}\mathbf{T}_{\bar{k},\mathbb{P}_m}\Pi_{\mathcal{Q}}^*|||_{\text{HS}}^2.$$

Without loss of generality, assume that the maximum $\max_m w_m |||\Pi_{\mathcal{Q}^\perp} \mathbf{T}_{\bar{k},\mathbb{P}_m} \Pi_{\mathcal{Q}}^*|||_{\text{HS}}^2$ is achieved at the index $m = 1$. Pick an orthonormal basis $\{h_m\}_{m=1}^\infty$ for $L^2(\bar{\mathbb{P}})$ with

$$\text{span}\{h_1, \ldots, h_M\} = \text{span}\{q_1, \ldots, q_M\}$$

and $h_1 = \tilde{q}_1$. By expanding the Hilbert–Schmidt norm in terms of this basis, we find

$$|||\Pi_{\mathcal{Q}^\perp} \mathbf{T}_{\bar{k},\mathbb{P}_1} \Pi_{\mathcal{Q}}^*|||_{\text{HS}}^2 \leq \left\|\mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1\right\|_{\bar{\mathbb{P}}}^2 - \left\langle \tilde{q}_1, \mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1 \right\rangle_{\bar{\mathbb{P}}}^2 + \sum_{m=2}^{M} \left\|\mathbf{T}_{\bar{k},\mathbb{P}_1} h_m\right\|_{\bar{\mathbb{P}}}^2. \tag{4.6.2}$$

By decomposing $\|\cdot\|_{\bar{\mathbb{P}}}^2$ according to the mixture representation of $\bar{\mathbb{P}}$, we obtain the inequality

$$\left\|\mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1\right\|_{\bar{\mathbb{P}}}^2 - \left\langle \tilde{q}_1, \mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1 \right\rangle_{\bar{\mathbb{P}}}^2 \leq w_1 \left\|\mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1\right\|_{\mathbb{P}_1}^2 - w_1^2 \left\langle \tilde{q}_1, \mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1 \right\rangle_{\mathbb{P}_1}^2 + \sum_{m=2}^{M} w_m \left\|\mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1\right\|_{\mathbb{P}_m}^2.$$

We claim that the following inequality holds:

$$\sum_{m=2}^{M} w_m \left\|\mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1\right\|_{\mathbb{P}_m}^2 \leq \frac{2(\mathcal{S}_{\text{max}}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}}))}{w_{\text{min}}^2}. \tag{4.6.3}$$

We take this claim as given for the moment, returning to prove it later.

Focusing on the expression $w_1 \left\|\mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1\right\|_{\mathbb{P}_1}^2 - w_1^2 \left\langle \tilde{q}_1, \mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1 \right\rangle_{\mathbb{P}_1}^2$, we add and subtract the term E_1, as defined in equation (4.6.1), and then apply triangle inequality, thereby finding that

$$\begin{aligned}
w_1 \left\|\mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1\right\|_{\mathbb{P}_1}^2 - w_1^2 \left\langle \tilde{q}_1, \mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1 \right\rangle_{\mathbb{P}_1}^2 = & \\
=& \frac{1}{w_1} \Big[\left\|(w_1 \mathbf{T}_{\bar{k},\mathbb{P}_1} + E_1 - E_1) \tilde{q}_1\right\|_{\mathbb{P}_1}^2 - w_1 \left\langle \tilde{q}_1, (w \mathbf{T}_{\bar{k},\mathbb{P}_1} + E_1 - E_1) \tilde{q}_1 \right\rangle_{\mathbb{P}_1}^2 \Big] \\
\leq& \frac{1}{w_1} \Big[\left(\|\tilde{q}_1\|_{\mathbb{P}_1} + |||E_1|||_{\text{op}} \|\tilde{q}_1\|_{\mathbb{P}_1} \right)^2 - w \left(\|\tilde{q}_1\|_{\mathbb{P}_1}^2 - \langle \tilde{q}_1, E_1 \tilde{q}_1 \rangle_{\mathbb{P}_1} \right)^2 \Big] \\
\leq& \frac{\|\tilde{q}_1\|_{\mathbb{P}_1}^2}{w_1} \Big[\frac{\sum_{m=2}^{M} w_m \|q_1\|_{\mathbb{P}_m}^2}{\|q_1\|_{\bar{\mathbb{P}}}^2} + (1 + w_1 \|\tilde{q}_1\|_{\mathbb{P}_1}^2)(2|||E_1|||_{\text{op}} + |||E_1|||_{\text{op}}^2) \Big].
\end{aligned}$$

Combining with the bound (4.6.3), we obtain the inequality

$$\left\|\mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1\right\|_{\bar{\mathbb{P}}}^2 - \left\langle \tilde{q}_1, \mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1 \right\rangle_{\bar{\mathbb{P}}}^2 \leq \frac{\mathcal{S}_{\text{max}}^2 + 2|||E_1|||_{\text{op}}}{w_{\text{min}}^2} + \frac{\mathcal{S}_{\text{max}}^2 + 8|||E_1|||_{\text{op}}}{w_{\text{min}}^2}.$$

Turning to the final term in equation (4.6.2), we first write

$$\left\|\mathbf{T}_{\bar{k},\mathbb{P}_1} h_m\right\|_{\bar{\mathbb{P}}}^2 = \int \Big(\int \sqrt{\bar{k}(x,y)} \sqrt{\bar{k}(x,y)} h_m(y) d\mathbb{P}_1(y) \Big)^2 d\bar{\mathbb{P}}(x).$$

By the Cauchy-Schwarz inequality, we have

$$\left\|\mathbf{T}_{\bar{k},\mathbb{P}_1} h_m\right\|_{\bar{\mathbb{P}}}^2 \leq \int \Big(\int \bar{k}(x,y) d\mathbb{P}_1(y) \Big) \Big(\int \bar{k}(x,y) h_m^2(y) d\mathbb{P}_1(y) \Big) d\bar{\mathbb{P}}(x)$$

$$\le \max_{\ell} \left\| \int k_\ell(x,y)\, d\mathbb{P}_\ell(y) \right\|_\infty^2 \|h_m\|_{\mathbb{P}_1}^2 .$$

Note that

$$\|h_m\|_{\mathbb{P}_1} \le \sup_{\substack{h\in \mathrm{span}\{q_2,\dots,q_M\} \\ \|h\|_{\bar{\mathbb{P}}}=1}} \|h\|_{\mathbb{P}_1} \le \sup_{a_\ell \in \mathbb{R}} \frac{\sum_{\ell=2}^{M} |a_\ell|\, \|q_\ell\|_{\mathbb{P}_1}}{\|\sum a_\ell q_\ell\|_{\bar{\mathbb{P}}}} = \max_{m\in\{2,\dots,M\}} \frac{\|q_m\|_{\mathbb{P}_1}}{\|q_m\|_{\bar{\mathbb{P}}}} \le \mathcal{S}_{\max}.$$

Therefore we conclude

$$|||\Pi_{\mathcal{Q}^\perp} \bar{\mathbf{T}} \Pi_{\mathcal{Q}}^*|||_{\mathrm{HS}}^2 \le \frac{M(12 + \|\tilde{p}_m\|_\infty^2)[\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})]}{w_{\min}^2}.$$

It remains to prove the bound (4.6.3). The fact $\left\|\bar{\mathbf{T}}\tilde{q}_m\right\|_{\bar{\mathbb{P}}} \le 1$ combined with the positivity of the kernel function implies that

$$1 \ge \left\|w_1 \mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1\right\|_{\bar{\mathbb{P}}}^2 = \sum_{m=1}^{M} w_m \left\|w_1 \mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1\right\|_{\mathbb{P}_m}^2 .$$

The term corresponding to $m=1$ in the expression above accounts for almost all of the sum. Indeed, if we examine the $m=1$ term, we find that

$$\begin{aligned} w_1 \left\|w_1 \mathbf{T}_{\bar{k},\mathbb{P}_1} \tilde{q}_1\right\|_{\mathbb{P}_1}^2 &= w_1 \left\|(w_1 \mathbf{T}_{\bar{k},\mathbb{P}_1} - \mathbf{T}_{k_1,\mathbb{P}_1})\tilde{q}_1 + \tilde{q}_1\right\|_{\mathbb{P}_1}^2 \\ &\ge w_1 \left\|\tilde{q}_1\right\|_{\mathbb{P}_1}^2 (1 - |||E_1|||_{\mathrm{op}})^2 \ge 1 - \mathcal{S}_{\max}^2 - 2|||E_1|||_{\mathrm{op}}, \end{aligned}$$

which completes the proof of the bound (4.6.3).

Proof of Lemma 2

Lower bound on $\sigma_{\min}(\mathbf{A})$: Let $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_M$ denote the nonzero eigenvalues of $\Pi_{\mathcal{Q}} \bar{\mathbf{T}} \Pi_{\mathcal{Q}}^*$. We seek a lower bound on $\tilde{\lambda}_M$. Since the spectrum of $\bar{\mathbf{T}}$ is contained in $[0,1]$, it suffices to lower bound $\sum_{m=1}^{M} \tilde{\lambda}_m^2 = |||\Pi_{\mathcal{Q}} \bar{\mathbf{T}} \Pi_{\mathcal{Q}}^*|||_{\mathrm{HS}}^2$. Let $\{h_1, \dots, h_M\}$ be an orthonormal basis of $\mathcal{Q}$ obtained from the Gram-Schmidt procedure on $\tilde{q}_1, \dots, \tilde{q}_M$. Note that

$$h_m = a_m \big(\tilde{q}_m - \sum_{\ell=1}^{m-1} \langle \tilde{q}_m, \tilde{q}_\ell \rangle_{\bar{\mathbb{P}}}\, \tilde{q}_\ell\big),$$

where $a_m := \left\|\tilde{q}_m - \sum_{\ell=1}^{m-1} \langle \tilde{q}_m, \tilde{q}_\ell \rangle_{\bar{\mathbb{P}}}\, \tilde{q}_\ell\right\|_{\bar{\mathbb{P}}}^{-1}$ denotes the normalizing constant.

By the definition of the Hilbert–Schmidt norm, we may write

$$|||\Pi_{\mathcal{Q}} \bar{\mathbf{T}} \Pi_{\mathcal{Q}}^*|||_{\mathrm{HS}}^2 \ge \sum_{m=1}^{M} \left\langle h_m, \bar{\mathbf{T}} h_m \right\rangle_{\bar{\mathbb{P}}}^2 . \tag{4.6.4}$$

By the loose bound $\left\langle \tilde{q}_m, \bar{\mathbf{T}} \tilde{q}_\ell \right\rangle_{\bar{\mathbb{P}}} \le 1$,

$$\left\langle h_m, \bar{\mathbf{T}} h_m \right\rangle_{\bar{\mathbb{P}}} \ge a_m^2 \big[\left\langle \tilde{q}_m, \bar{\mathbf{T}} \tilde{q}_m \right\rangle_{\bar{\mathbb{P}}} - 2M \big(\max_{m \ne \ell} \langle \tilde{q}_m, \tilde{q}_\ell \rangle_{\bar{\mathbb{P}}} \big) \big].$$

If we decompose the Laplacian operator via $\bar{\mathbf{T}} = \sum_{m=1}^{M} w_m \mathbf{T}_{\bar{k},\mathbb{P}_m}$, and if also write out the inner product $\langle \cdot, \cdot \rangle_{\bar{\mathbb{P}}}$ as a mixture of inner products, we find that

$$\begin{aligned}\left\langle \tilde{q}_m, \bar{\mathbf{T}} \tilde{q}_m \right\rangle_{\bar{\mathbb{P}}} &\geq \left\langle \tilde{q}_m, w_m \mathbf{T}_{\bar{k},\mathbb{P}_m} \tilde{q}_m \right\rangle_{\bar{\mathbb{P}}} \geq w_m \left\langle \tilde{q}_m, w_m \mathbf{T}_{\bar{k},\mathbb{P}_m} \tilde{q}_m \right\rangle_{\mathbb{P}_m} \\ &\geq w_m \big(\langle \tilde{q}_m, \mathbf{T}_{k_m,\mathbb{P}_m} \tilde{q}_m \rangle_{\mathbb{P}_m} - |||E_m|||_{\text{op}} \, \|\tilde{q}_m\|_{\mathbb{P}_m}^2 \big) \\ &= w_m \, \|\tilde{q}_m\|_{\mathbb{P}_m}^2 \, \big(1 - |||E_m|||_{\text{op}}\big).\end{aligned}$$

Plugging this lower bound into equation (4.6.4), we obtain

$$M \geq |||\Pi_{\mathcal{Q}} \bar{\mathbf{T}} \Pi_{\mathcal{Q}}^*|||_{\text{HS}}^2 \geq \sum_{m=1}^{M} a_m^4 \big[(1 - \mathcal{S}_{\text{max}}^2)(1 - |||E_m|||_{\text{op}}) - 6M\mathcal{S}_{\text{max}}\big]^2.$$

To finish the argument, we must lower bound a_m. Note that

$$\left\| \tilde{q}_m - \sum_{\ell=1}^{m-1} \langle \tilde{q}_m, \tilde{q}_\ell \rangle_{\bar{\mathbb{P}}} \, \tilde{q}_\ell \right\|_{\bar{\mathbb{P}}} \leq \big(1 + (M-1)(2\mathcal{S}_{\text{max}} + \mathcal{S}_{\text{max}}^2)\big) \leq (1 + 3M\mathcal{S}_{\text{max}}).$$

When $3M\mathcal{S}_{\text{max}} < 1$, as is implied by the hypothesis of Theorem 1, some further algebra yields

$$\tilde{\lambda}_M \geq 1 - 13M[\mathcal{S}_{\text{max}}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})]^{1/2} \geq 1 - 13\frac{[\mathcal{S}_{\text{max}}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})]^{1/2}}{w_{\text{min}}}.$$

Upper bound on $\sigma_{\text{max}}(\mathbf{B})$: We pursue an upper bound on $|||\Pi_{\mathcal{Q}^\perp} \bar{\mathbf{T}} \Pi_{\mathcal{Q}^\perp}^*|||_{\text{op}} \leq |||\bar{\mathbf{T}} \Pi_{\mathcal{Q}^\perp}^*|||_{\text{op}}$. By the definition of operator norm and the decomposability of $\|\|_{\bar{\mathbb{P}}}$ via mixture components,

$$|||\bar{\mathbf{T}} \Pi_{\mathcal{Q}^\perp}^*|||_{\text{op}} = \sup_{\substack{f \in \mathcal{Q}^\perp \\ \|f\|_{\bar{\mathbb{P}}} \leq 1}} \left\| \bar{\mathbf{T}} f \right\|_{\bar{\mathbb{P}}} \leq \sup_{\substack{f \in \mathcal{Q}^\perp \\ \|f\|_{\bar{\mathbb{P}}} \leq 1}} \max_m \left\| \bar{\mathbf{T}} f \right\|_{\mathbb{P}_m} \leq \max_m \sup_{\substack{f \in \mathcal{Q}^\perp \\ \|f\|_{\bar{\mathbb{P}}} \leq 1}} \left\| \bar{\mathbf{T}} f \right\|_{\mathbb{P}_m}.$$

We upper bound the supremum in the right hand side above for each $m = 1, \ldots, M$. Consider an arbitrary feasible f (depending on m), and define $g = f - \langle f, \widehat{q_m} \rangle_{\mathbb{P}_m} \widehat{q_m}$, where $\widehat{q_m} = \frac{q_m}{\|q_m\|_{\mathbb{P}_m}}$. Since $f \in \mathcal{Q}^\perp$, we expect $\|g - f\|_{\bar{\mathbb{P}}}$ to be small. Peeling off small terms, we obtain the inequality

$$\left\| \bar{\mathbf{T}} f \right\|_{\mathbb{P}_m} \leq \|\mathbf{T}_{k_m,\mathbb{P}_m} g\|_{\mathbb{P}_m} + \|g - f\|_{\mathbb{P}_m} + \sum_{\ell \neq m} w_\ell \left\| \mathbf{T}_{\bar{k},\mathbb{P}_\ell} f \right\|_{\mathbb{P}_m} + |||E_m|||_{\text{op}} \|f\|_{\mathbb{P}_m}. \tag{4.6.5}$$

The three terms besides $\|\mathbf{T}_{k_m,\mathbb{P}_m} g\|_{\mathbb{P}_m}$ on the right hand side of equation (4.6.5) are small when the overlap between mixture components is small. Indeed, the first of these is

$$\begin{aligned}\|g - f\|_{\bar{\mathbb{P}}} = \|\widehat{q_m}\|_{\bar{\mathbb{P}}} \left| \langle f, \widehat{q_m} \rangle_{\mathbb{P}_m} \right| &\overset{(i)}{=} \frac{\|q_m\|_{\bar{\mathbb{P}}}}{w_m \|q_m\|_{\mathbb{P}_m}} \left| \sum_{\ell \neq m} w_\ell \langle f, \widehat{q_m} \rangle_{\mathbb{P}_\ell} \right| \\ &= \frac{\|q_m\|_{\bar{\mathbb{P}}}^2}{w_m \, \|q_m\|_{\mathbb{P}_m}^2} \left| \sum_{\ell \neq m} w_\ell \langle f, \tilde{q}_m \rangle_{\mathbb{P}_\ell} \right| \leq \Big(1 + \frac{\mathcal{S}_{\text{max}}^2}{w_{\text{min}}}\Big) \left| \sum_{\ell \neq m} w_\ell \, \|f\|_{\mathbb{P}_\ell} \right| \mathcal{S}_{\text{max}} \\ &\overset{(ii)}{\leq} 2\mathcal{S}_{\text{max}}\end{aligned}$$

Here equality (i) follows from the identity $0 = \langle f, q_m \rangle_{\bar{\mathbb{P}}} = \sum_{\ell=1}^{M} w_\ell \langle f, q_m \rangle_{\mathbb{P}_\ell}$, whereas inequality (ii) follows from Jensen's inequality, the bound $\|f\|_{\bar{\mathbb{P}}} \leq 1$, and the assumption $\mathcal{S}_{\max}^2 \leq w_{\min}$.

The second of these small terms is $\sum_{\ell \neq m} \left\| w_\ell \mathbf{T}_{\bar{k}, \mathbb{P}_\ell} f \right\|_{\mathbb{P}_m}$. By Jensen's inequality and inequality (4.6.3), we obtain

$$\sum_{\ell \neq m} \left\| w_\ell \mathbf{T}_{\bar{k}, \mathbb{P}_\ell} f \right\|_{\mathbb{P}_m} \leq \sqrt{\sum_{\ell \neq m} w_\ell \left\| \mathbf{T}_{\bar{k}, \mathbb{P}_\ell} f \right\|_{\mathbb{P}_m}^2} \leq \frac{\sqrt{2}[\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})]^{1/2}}{w_{\min}}.$$

To complete the argument, we bound $\|\mathbf{T}_{k_m, \mathbb{P}_m} g\|_{\mathbb{P}_m}$ by the second eigenvalue of $\mathbf{T}_{k_m, \mathbb{P}_m}$, which we denote by $\lambda_2(\mathbf{T}_{k_m, \mathbb{P}_m})$. Indeed, note that

$$\|\mathbf{T}_{k_m, \mathbb{P}_m} g\|_{\mathbb{P}_m} \leq \sup_{\substack{\|h\|_{\mathbb{P}_m} \leq 1 \\ \langle h, q_m \rangle_{\mathbb{P}_m} = 0}} \|\mathbf{T}_{k_m, \mathbb{P}_m} h\|_{\mathbb{P}_m} = \lambda_2(\mathbf{T}_{k_m, \mathbb{P}_m}).$$

To control $\lambda_2(\mathbf{T}_{k_m, \mathbb{P}_m})$, we use a version of Cheeger's inequality (see Appendix 4.8 for the theorem statement). Consider the transition kernel $\Pi(x, dy) := \frac{k(x,y) d\mathbb{P}_m(y)}{p(x)}$, where $p(x) = \int k(x, y) d\mathbb{P}_m(y)$. Let $Z = \int_{\mathcal{X}} p(x) d\mathbb{P}_m(x)$, and let $\tilde{\mathbb{P}}_m$ denote the probability measure on $\mathcal{X}$ that assigns to a measurable set $S \subset \mathcal{X}$ the mass

$$\tilde{\mathbb{P}}_m(S) = \frac{1}{Z} \int_{y \in S} p(y) d\mathbb{P}_m(y).$$

It is easy to verify that $\tilde{\mathbb{P}}_m$ is the invariant probability measure for the transition kernel Π. The transition kernel Π induces a linear transformation $\mathbf{T}_{\text{Asym}} : L^2(\tilde{\mathbb{P}}_m) \to L^2(\tilde{\mathbb{P}}_m)$ via

$$(\mathbf{T}_{\text{Asym}} f)(x) = \int \frac{k(x, y)}{p(x)} f(y) d\mathbb{P}_m(y).$$

Cheeger's inequality bounds the second eigenvalue of this transformation. In particular, by Theorem 4.8.3, we have

$$1 - \frac{\Gamma(\Pi)^2}{8} \geq \lambda_2(\mathbf{T}_{\text{Asym}}), \quad \text{where } \Gamma(\Pi) := \inf_{0 < \tilde{\mathbb{P}}_m(S) < 1} \frac{\int_S \Pi(x, S^c) \tilde{\mathbb{P}}_m(dx)}{\tilde{\mathbb{P}}_m(S) \tilde{\mathbb{P}}_m(S^c)}.$$

However, our choice of Π implies that $\Gamma(\Pi) \leq \Gamma$. It is straight forward to verify that the spectrum of the asymmetric Laplacian $\mathbf{T}_{\text{Asym}}$ coincides with the spectrum of the symmetric Laplacian $\mathbf{T}_{k_m, \mathbb{P}_m}$. Therefore

$$1 - \frac{\Gamma^2}{8} \geq \lambda_2(\mathbf{T}_{k_m, \mathbb{P}_m}).$$

Finally, inequality (4.6.5) reduces to the inequality

$$\|\bar{\mathbf{T}} \Pi_{\mathcal{Q}^\perp}^* \|_{\text{op}} \leq 1 - \frac{\Gamma^2}{8} + \frac{3[\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})]^{1/2}}{w_{\min}}.$$

4.7 Supplementary proofs for Theorem 2

This appendix is devoted to the second step in proof of Theorem 2 and the proofs of Lemmas 3 and 4. Note that we present the proof of Lemma 4 before the proof of Lemma 3 because the two proofs mostly follow the same steps and the proof of Lemma 4 is slightly easier.

Proof of Step 2

Recall that a diverse M-tuple is good if all its distinct pairs form angles within θ of $\frac{\pi}{2}$ or $\frac{3\pi}{2}$. Otherwise we call it bad. In Step 1, we showed that a uniformly chosen diverse M-tuple is good with probability $1-p$. Since there are $\prod_{i=1}^{M} n_i$ diverse M-tuples, at most $\lceil p(\prod_{i=1}^{M} n_i)\rceil$ of them can be bad.

Let $e_1 \in \mathbb{R}^M$ be the point in the fewest bad M-tuples. Without loss of generality, we may assume that e_1 has latent label 1. By the minimality of e_1, it is contained in at most $p\prod_{i=2}^{M} n_i$ bad M-tuples. Let $G_1 \subset \mathbb{R}^{M\times M}$ denote the set of good M-tuples containing e_1. Let $e_2 \in \mathbb{R}^M$ be the most frequent point, different from e_1, occurring in the M-tuples in G_1. Suppose without loss of generality that e_2 has latent label 2. Let G_2 denote the set of diverse M-tuples in G_1 with e_2 in the second coordinate. By the minimality of e_2, we find that

$$|G_2| \geq (1-p)n_3 \dots n_M.$$

Continuing in this way, choose points $e_3, \dots, e_{M-1}$ and construct $G_3, \dots, G_{M-1}$. Note that

$$G_{M-1} = \{(e_1, \dots, e_{M-1}, z) : z \text{ has label } M \text{ and } (e_1, \dots, e_{M-1}, z) \text{ is good}\}.$$

By the minimality of $e_1, \dots, e_{M-1}$, we find that $|G_{M-1}| \geq (1-p)n_M$. Moreover all the elements of G_{M-1} are good M-tuples. However, all the M-tuples of G_{M-1} contain $e_1, \dots, e_{M-1}$.

Define the set

$$Z := \big\{z \mid (e_1, \dots, e_{M-1}, z) \in G_{M-1}\big\}.$$

Let G_Z be the set of all good M-tuples with one coordinate in Z. Suppose a fraction p' of the diverse M-tuples $(y_1, \dots, y_{M-1}, z) \in G_Z$ are incompatible with e_1 in the sense that $(e_1, y_2, \dots, y_{M-1}, z)$ is bad. This implies that e_1 is in at least $p'\,|G_Z|$ bad M-tuples. By the minimality of e_1, we have the inequality

$$p' \leq \frac{p}{1-p}.$$

Let $G_Z^1 \subset G_Z$ be the subset of G_Z that is compatible with e_1. The bound on p' implies that

$$\left|G_Z^1\right| \geq \frac{1-2p}{1-p}\left|G_Z\right|.$$

Define $G_Z^2 \subset G_Z^1$ to be the subset of M-tuples in G_Z^1 compatible with e_2. By the choice of e_2,

$$\left|G_Z^2\right| \geq \frac{1-2p}{1-p}\left|G_Z^1\right|.$$

Continuing in this way, define $G_Z^{M-1} \subset \cdots \subset G_Z^2 \subset G_Z^1$ with

$$\frac{\left|G_Z^{M-1}\right|}{n_1 \dots n_M} \geq \left[1-2p\right]\left[\frac{1-2p}{1-p}\right]^{M-2} \geq 1-Mp.$$

By construction, all of the elements of G_Z^{M-1} are good. Moreover, all are compatible with $e_1, \dots, e_{M-1}$. This completes the proof of Step 2.

Proof of Lemma 4

Recall that $\mathcal{G}$ denotes the principal eigenspace of $\bar{\mathbf{H}}$ and $\tilde{\mathcal{G}}$ denotes the principal eigenspace of $\tilde{\mathbf{H}}$. We apply a result due to Stewart [107], stated precisely as Theorem 4.8.2 in Appendix 4.8, to show that $\mathcal{G}$ is an approximate invariant subspace of $\tilde{\mathbf{H}}$. The spectral separation of two operators A and B is given by $\operatorname{sep}(A,B) := \inf\{|a-b| : a \in \sigma(A), b \in \sigma(B)\}$. By Theorem 4.8.2, if

$$\frac{\|\Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*\|_{\text{HS}}}{\operatorname{sep}(\Pi_{\mathcal{G}}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*, \Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}^\perp}^*)} \leq \frac{1}{2},$$

then there is an operator $A : \mathcal{G} \to \mathcal{G}^\perp$ satisfying

$$\|A\|_{\text{HS}} \leq \frac{2\|\Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*\|_{\text{HS}}}{\operatorname{sep}(\Pi_{\mathcal{G}}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*, \Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}^\perp}^*)}$$

such that $\operatorname{Range}(\Pi_{\mathcal{G}}^* - \Pi_{\mathcal{G}^\perp}^* A)$ is an invariant subspace of $\tilde{\mathbf{H}}$. By the argument of Theorem 1, we obtain

$$\rho(\mathcal{G}, \tilde{\mathcal{G}}) \leq \frac{2\|\Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*\|_{\text{HS}}}{\operatorname{sep}(\Pi_{\mathcal{G}}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*, \Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}^\perp}^*)}.$$

We complete the proof of the Lemma in two steps; we first upper bound $\|\Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*\|_{\text{HS}}$, and then lower bound the eigengap term $\operatorname{sep}(\Pi_{\mathcal{G}}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*, \Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}^\perp}^*)$.

Step 1: In Step 1, we control the tails of the random variable $\|\Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*\|_{\text{HS}}$. We work instead with the operator norm, and this causes us to pick up an additional factor of $\sqrt{M}$ because

$$\|\Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*\|_{\text{HS}} \leq \sqrt{M}\|\Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*\|_{\text{op}}.$$

Since $\Pi_{\mathcal{G}^\perp}\bar{\mathbf{H}}\Pi_{\mathcal{G}}^* = 0$, we are free to add this term inside the operator norm, which yields

$$\|\Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*\|_{\text{op}} \leq \|\tilde{\mathbf{H}} - \bar{\mathbf{H}}\|_{\text{op}}.$$

We complete the first step by proving that the random variable $\|\bar{\mathbf{H}} - \tilde{\mathbf{H}}\|_{\text{op}}$ has subgaussian tails.

By the definition of operator norm,

$$\|\bar{\mathbf{H}} - \tilde{\mathbf{H}}\|_{\text{op}} = \sup_{\|h\|_{\bar{\mathcal{H}}} \leq 1} \left| \left\langle h(y), \int \bar{k}(x,y)h(x)[d\bar{\mathbb{P}}_n(x) - d\bar{\mathbb{P}}(x)] \right\rangle_{\bar{\mathcal{H}}} \right|.$$

By exchanging the order of inner product and integration and using the representer property, we find that

$$\|\bar{\mathbf{H}} - \tilde{\mathbf{H}}\|_{\text{op}} = \sup_{\|h\|_{\bar{\mathcal{H}}} \leq 1} \left| \frac{1}{n} \sum_{i=1}^{n} h^2(X_i) - \int h^2(x) d\bar{\mathbb{P}}(x) \right|.$$

The object on the right hand side is the supremum of the empirical process over the function class $\mathcal{F}^2 = \{h^2 : \|h\|_{\bar{\mathcal{H}}} \leq 1\}$. By definition $\bar{k}(x, y) \leq \frac{b}{r^2}$ and therefore the function class $\mathcal{F}^2$ is uniformly bounded in the sense that for any h with $\|h\|_{\bar{\mathcal{H}}} \leq 1$, $h^2(x) = \langle h, \bar{k}_x \rangle^2_{\bar{\mathcal{H}}} \leq \|\bar{k}_x\|^2_{\bar{\mathcal{H}}} \leq \frac{b}{r^2}$. We therefore may apply a standard concentration results for empirical processes over bounded function classes [13] to obtain

$$\|\bar{\mathbf{H}} - \tilde{\mathbf{H}}\|_{\text{op}} \leq 2\mathcal{R}_n(\mathcal{F}^2) + \delta_0, \tag{4.7.1}$$

with probability at least $1 - 2\exp\left(-\frac{n{\delta_0}^2 r^4}{8b^2}\right)$. By standard results on Rademacher and Gaussian complexity [12], we have $\sqrt{\frac{2}{\pi}}\mathcal{R}_n(\mathcal{F}^2) \leq \frac{2b}{r^2}\sqrt{\frac{\mathbb{E}\bar{k}(\bar{X},\bar{X})}{n}}$.

Step 2: To complete the proof of the lemma, we must control the eigengap term. By Stewart [107, Thm 2.3], the spectral separation sep is stable to perturbations in its arguments. Adding and subtracting $\Pi_{\mathcal{G}}\bar{\mathbf{H}}\Pi^*_{\mathcal{G}}$ in the first argument of sep below (and $\Pi_{\mathcal{G}^\perp}\bar{\mathbf{H}}\Pi^*_{\mathcal{G}^\perp}$ in the second), we find that

$$\text{sep}(\Pi_{\mathcal{G}}\tilde{\mathbf{H}}\Pi^*_{\mathcal{G}}, \Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi^*_{\mathcal{G}^\perp}) \geq \text{sep}(\Pi_{\mathcal{G}}\bar{\mathbf{H}}\Pi^*_{\mathcal{G}}, \Pi_{\mathcal{G}^\perp}\bar{\mathbf{H}}\Pi^*_{\mathcal{G}^\perp}) - 2\|\bar{\mathbf{H}} - \tilde{\mathbf{H}}\|_{\text{op}}.$$

Note that $\text{sep}(\Pi_{\mathcal{G}}\bar{\mathbf{H}}\Pi^*_{\mathcal{G}}, \Pi_{\mathcal{G}^\perp}\bar{\mathbf{H}}\Pi^*_{\mathcal{G}^\perp}) = \text{sep}(\Pi_{\mathcal{R}}\bar{\mathbf{T}}\Pi^*_{\mathcal{R}}, \Pi_{\mathcal{R}^\perp}\bar{\mathbf{T}}\Pi^*_{\mathcal{R}^\perp})$, and we can replace the projections onto $\mathcal{R}$ by projections onto $\mathcal{Q}$ using the same trick. This is helpful since we know from our population level analysis that

$$\text{sep}(\Pi_{\mathcal{Q}}\bar{\mathbf{T}}\Pi^*_{\mathcal{Q}}, \Pi_{\mathcal{Q}^\perp}\bar{\mathbf{T}}\Pi^*_{\mathcal{Q}^\perp}) \geq \frac{\Gamma^2}{16},$$

and therefore (using $\rho(\mathcal{R}, \mathcal{Q}) \leq 1$), we obtain

$$\text{sep}(\Pi_{\mathcal{G}}\tilde{\mathbf{H}}\Pi^*_{\mathcal{G}}, \Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi^*_{\mathcal{G}^\perp}) \geq \frac{\Gamma^2}{16} - 6\rho(\mathcal{R}, \mathcal{Q}) - 2\|\bar{\mathbf{H}} - \tilde{\mathbf{H}}\|_{\text{op}}. \tag{4.7.2}$$

On the set $\mathcal{A}_0 = \{6\rho(\mathcal{R}, \mathcal{Q}) + 2\|\bar{\mathbf{H}} - \tilde{\mathbf{H}}\|_{\text{op}} \leq \frac{\Gamma^2}{32}\}$ we obtain the lower bound

$$\text{sep}(\Pi_{\mathcal{G}}\tilde{\mathbf{H}}\Pi^*_{\mathcal{G}}, \Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi^*_{\mathcal{G}^\perp}) \geq \frac{\Gamma^2}{32}. \tag{4.7.3}$$

By equation (4.7.1), the set $\mathcal{A}_0$ has probability at least $1 - 2e^{\frac{-n{\delta_0}^2 r^4}{8b^2}}$. Finally, combining the results of Step 1 and Step 2, we obtain the inequality

$$\rho(\mathcal{G}, \tilde{\mathcal{G}}) \leq \frac{64\sqrt{M}}{\Gamma^2}\left(\frac{2\sqrt{2\pi}b\sqrt{\mathbb{E}\bar{k}(X, X)}}{r^2\sqrt{n}} + \delta_0\right),$$

with probability at least $1 - 2e^{\frac{-n{\delta_0}^2 r^4}{8b^2}}$. Note that we write the statement of the lemma in terms of $\delta = \frac{r^2\delta_0}{2\sqrt{2\pi}b\sqrt{\mathbb{E}\bar{k}(X,X)}}$.

Proof of Lemma 3

Recall that $\mathcal{V}$ denotes the principal eigenspace of $\hat{\mathbf{T}}$, and $\tilde{\mathcal{V}}$ denotes the principal eigenspace of $\tilde{\mathbf{T}}$. We apply Stewart's operator perturbation theorem (see section 4.8) to show that $\tilde{\mathcal{V}}$ is an approximate invariant subspace of $\hat{\mathbf{T}}$. By Stewart [107, Thm 3.6], if

$$\frac{\|\Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}}\|_{\mathrm{HS}}}{\mathrm{sep}(\Pi_{\tilde{\mathcal{V}}}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}}, \Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}^\perp})} \le \frac{1}{2},$$

then there is an operator $A : \tilde{\mathcal{V}} \to \tilde{\mathcal{V}}^\perp$ satisfying

$$\|A\|_{\mathrm{HS}} \le \frac{2\|\Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}}\|_{\mathrm{HS}}}{\mathrm{sep}(\Pi_{\tilde{\mathcal{V}}}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}}, \Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}^\perp})}$$

such that $\mathrm{Range}(\Pi^*_{\tilde{\mathcal{V}}} - \Pi^*_{\tilde{\mathcal{V}}^\perp}A)$ is an invariant subspace of $\hat{\mathbf{T}}$. By the argument of Theorem 1, we obtain

$$\rho(\mathcal{V}, \tilde{\mathcal{V}}) \le \frac{2\|\Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}}\|_{\mathrm{HS}}}{\mathrm{sep}(\Pi_{\tilde{\mathcal{V}}}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}}, \Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}^\perp})}. \tag{4.7.4}$$

We complete the proof of the lemma in two steps; we first upper bound $\|\Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}}\|_{\mathrm{HS}}$, and then lower bound the eigengap term $\mathrm{sep}(\Pi_{\tilde{\mathcal{V}}}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}}, \Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}^\perp})$.

Step 1: In Step 1, we show that the random variable $\|\Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}}\|_{\mathrm{HS}}$ has sub-Gaussian tails. Note that

$$\|\Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi^*_{\tilde{\mathcal{V}}}\|_{\mathrm{HS}} \le \sqrt{M}\|\tilde{\mathbf{T}} - \hat{\mathbf{T}}\|_{\mathrm{op}}.$$

Recall that the kernels of these two integral operators are given by

$$\bar{k}(x,y) = \frac{k(x,y)}{\bar{q}(x)\bar{q}(y)}, \quad \bar{k}^n(x,y) = \frac{k(x,y)}{\bar{q}_n(x)\bar{q}_n(y)}.$$

We show that the kernels quite similar for large n by controlling the difference $\|\bar{q}^2 - \bar{q}_n^2\|_\infty$. Rewrite the difference as

$$\left\|\bar{q}^2 - \bar{q}_n^2\right\|_\infty = \sup_x \left|\mathbb{E}k(x,X) - \frac{1}{n}\sum_{i=1}^n k(x,X_i)\right|.$$

Note that the right hand side above is the supremum of the empirical process over the class of functions $\mathcal{F} = \{k(x,\cdot) : x \in \mathbb{X}\}$. This function class is uniformly bounded in the sense that $k(x,y) \le b$. Therefore, by a standard concentration result for empirical processes [13], we have

$$\sup_x \left|\bar{\mathbb{P}}k(x,\cdot) - \bar{\mathbb{P}}_n k(x,\cdot)\right| \le 2\mathcal{R}_n\left\{k(x,\cdot) : x \in \mathbb{X}\right\} + \delta_1, \tag{4.7.5}$$

with probability at least $1 - 2e^{-\frac{n{\delta_1}^2}{8b^2}}$. The Rademacher complexity is upper bounded by

$$\mathcal{R}_n(\mathcal{F}) \le \sqrt{\frac{\pi b}{2n}}\sqrt{\mathbb{E}k(\bar{X},\bar{X})}.$$

Let $Q, Q_n : L^2(\bar{\mathbb{P}}_n) \to L^2(\bar{\mathbb{P}}_n)$ denote operators corresponding to pointwise multiplication by $\bar{q}$ and $\bar{q}_n$, respectively. In other words, for $f \in L^2(\bar{\mathbb{P}}_n)$, we have

$$Qf = \bar{q}f, \quad Q_n f = \bar{q}_n f.$$

Recall we assume there is a scalar $r > 0$ such that $\bar{\mathbb{P}}\{\bar{q}(X) < r\} = 0$. Therefore, the operator Q is invertible with bounded inverse Q^{-1} given by pointwise multiplication by $\frac{1}{\bar{q}}$. In particular, we have the bound $\|Q^{-1}\|_{\mathrm{op}} \leq \frac{1}{r}$. Let $r_n = \min_i \bar{q}_n(X_i)$.

The advantage of having introduced the operators Q and Q_n is that we can now write

$$\tilde{\mathbf{T}} = Q^{-1}\mathbf{T}_{k,\bar{\mathbb{P}}_n}Q^{-1}, \quad \hat{\mathbf{T}} = Q_n^{-1}\mathbf{T}_{k,\bar{\mathbb{P}}_n}Q_n^{-1}.$$

By the triangle inequality for operator norm, we have

$$\|\tilde{\mathbf{T}} - \hat{\mathbf{T}}\|_{\mathrm{op}} \leq \|\mathbf{T}_{k,\bar{\mathbb{P}}_n}\|_{\mathrm{op}}\big(\|Q^{-1} - Q_n^{-1}\|_{\mathrm{op}}^2 + 2\|Q^{-1} - Q_n^{-1}\|_{\mathrm{op}}\|Q_n^{-1}\|_{\mathrm{op}}\big). \tag{4.7.6}$$

We can control the finite sample error in Q with our empirical process bound (4.7.5). Note in particular that

$$\|Q^{-1} - Q_n^{-1}\|_{\mathrm{op}} \leq \frac{\|\bar{q}_n - \bar{q}\|_\infty}{rr_n} \leq \frac{1}{r^2 r_n}\big\|\bar{q}^2(x) - \bar{q}_n^2(x)\big\|_\infty. \tag{4.7.7}$$

As is the case for any kernel integral operator, we have $\|\mathbf{T}_{k,\bar{\mathbb{P}}}\|_{\mathrm{op}} \leq \|k\|_{\bar{\mathbb{P}}}$. However, equation (4.7.6) involves the empirical version, $\|\mathbf{T}_{k,\bar{\mathbb{P}}_n}\|_{\mathrm{op}}$. The concentration of $\mathbf{T}_{k,\bar{\mathbb{P}}_n}$ about $\mathbf{T}_{k,\bar{\mathbb{P}}}$ follows from Rosasco et al. [96, Thm. 7], which implies that $\|\mathbf{T}_{k,\bar{\mathbb{P}}_n}\|_{\mathrm{op}} \leq 2\|\mathbf{T}_{k,\bar{\mathbb{P}}}\|_{\mathrm{op}}$ with probability at least $1 - 2e^{\frac{-n\|k\|_{\bar{\mathbb{P}}}^2}{8b^2}}$. Therefore $\|\mathbf{T}_{k,\bar{\mathbb{P}}_n}\|_{\mathrm{op}} \leq 2\,\|k\|_{\bar{\mathbb{P}}}$ with the same probability. Using this fact, after some algebra equation (4.7.6) reduces to

$$\|\tilde{\mathbf{T}} - \hat{\mathbf{T}}\|_{\mathrm{op}} \leq \frac{16\,\|k\|_{\bar{\mathbb{P}}}}{r^4}\big(\delta_1 + \frac{\sqrt{2\pi}b}{\sqrt{n}}\big), \tag{4.7.8}$$

with probability at least $1 - 4e^{\frac{-n\delta_1{}^2}{8b^2}}$ whenever $\delta_1 + \frac{\sqrt{2\pi}b}{\sqrt{n}} \leq \frac{r^2}{2}$, and $\delta_1 \leq \|k\|_{\bar{\mathbb{P}}}$.

Step 2: We lower bound the eigengap term $\mathrm{sep}(\Pi_{\tilde{\mathcal{V}}}\hat{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}}^*, \Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}^\perp}^*)$. By Stewart [107, Thm 2.3], the spectral separation sep is stable to perturbations in its arguments. Adding and subtracting $\Pi_{\tilde{\mathcal{V}}}\tilde{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}}^*$ in the first argument of sep below (and $\Pi_{\tilde{\mathcal{V}}^\perp}\tilde{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}^\perp}^*$ in the second) yields

$$\mathrm{sep}(\Pi_{\tilde{\mathcal{V}}}\hat{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}}^*, \Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}^\perp}^*) \geq \mathrm{sep}(\Pi_{\tilde{\mathcal{V}}}\tilde{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}}^*, \Pi_{\tilde{\mathcal{V}}^\perp}\tilde{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}^\perp}^*) - 2\|\hat{\mathbf{T}} - \tilde{\mathbf{T}}\|_{\mathrm{op}}.$$

Note that $\mathrm{sep}(\Pi_{\tilde{\mathcal{V}}}\tilde{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}}^*, \Pi_{\tilde{\mathcal{V}}^\perp}\tilde{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}^\perp}^*) = \mathrm{sep}(\Pi_{\tilde{\mathcal{G}}}\tilde{\mathbf{H}}\Pi_{\tilde{\mathcal{G}}}^*, \Pi_{\tilde{\mathcal{G}}^\perp}\tilde{\mathbf{H}}\Pi_{\tilde{\mathcal{G}}^\perp}^*)$, and we can replace projections onto $\tilde{\mathcal{G}}$ by projections onto $\mathcal{G}$ by again leveraging the stability of sep to perturbations in its arguments. In this way we obtain (using $\rho(\mathcal{G}, \tilde{\mathcal{G}}) \leq 1$)

$$\begin{aligned}\mathrm{sep}(\Pi_{\tilde{\mathcal{V}}}\hat{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}}^*, \Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}^\perp}^*) \geq\, &\mathrm{sep}(\Pi_{\mathcal{G}}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*, \Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}^\perp}^*) - 2\|\hat{\mathbf{T}} - \tilde{\mathbf{T}}\|_{\mathrm{op}}\\ &- 6\|\tilde{\mathbf{H}}\|_{\mathrm{op}}\rho(\mathcal{G}, \tilde{\mathcal{G}}).\end{aligned}$$

Recall from equation (4.7.2) that $\operatorname{sep}(\Pi_{\mathcal{G}}\tilde{\mathbf{H}}\Pi_{\mathcal{G}}^*, \Pi_{\mathcal{G}^\perp}\tilde{\mathbf{H}}\Pi_{\mathcal{G}^\perp}^*) \geq \frac{\Gamma^2}{32}$ on $\mathcal{A}_0$. Therefore, we obtain the lower bound

$$\operatorname{sep}(\Pi_{\tilde{\mathcal{V}}}\hat{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}}^*, \Pi_{\tilde{\mathcal{V}}^\perp}\hat{\mathbf{T}}\Pi_{\tilde{\mathcal{V}}^\perp}^*) \geq \frac{\Gamma^2}{64},$$

valid on the set $\mathcal{A}_0 \cap \mathcal{A}_1$, where $\mathcal{A}_1 = \{2\|\hat{\mathbf{T}} - \tilde{\mathbf{T}}\|_{\text{op}} + 6\|\tilde{\mathbf{H}}\|_{\text{op}}\rho(\mathcal{G}, \tilde{\mathcal{G}}) \leq \frac{\Gamma^2}{64}\}$. By equation (4.7.8), the set $\mathcal{A}_1$ has probability at least $1 - 4e^{\frac{-n\delta_1{}^2}{8b^2}}$. Recall that $\mathcal{A}_0$ has probability at least $1 - 2e^{\frac{-n\delta_0{}^2r^4}{8b^2}}$. Therefore $\mathcal{A}_0 \cap \mathcal{A}_1$ has probability at least $1 - 4e^{\frac{-n\delta_1{}^2}{8b^2}} - 2e^{\frac{-n\delta_0{}^2r^4}{8b^2}}$. This completes Step 2.

Combining the results of Step 1 and Step 2, inequality (4.7.4) simplifies to

$$\rho(\mathcal{V}, \tilde{\mathcal{V}}) \leq \frac{512\sqrt{M}\,\|k\|_{\bar{\mathbb{P}}}}{r^4\Gamma^2}\big(\frac{\sqrt{2\pi}b}{\sqrt{n}} + \delta_1\big), \tag{4.7.9}$$

with probability at least $1 - 4e^{\frac{-n\delta_1{}^2}{8b^2}} - 2e^{\frac{-n\delta_0{}^2r^4}{8b^2}}$. In the statement of the lemma we write $\delta_1 = \sqrt{2\pi}b\delta$, and $\delta_0 = \frac{2\sqrt{2\pi}b\sqrt{\mathbb{E}\bar{k}(X,X)}}{r^2}\delta$.

4.8 Background

In this appendix, we provide some background on kernel integral operators, reproducing kernel Hilbert spaces, and operator perturbation theory.

Kernel integral operators

Given a probability measure $\mathbb{P}$ on a measurable space $\mathcal{X}$, and a $\mathbb{P}$-square integral kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the kernel integral operator on $L^2(\mathbb{P}, \mathcal{X})$ is the linear operator

$$\mathbf{T}_{k,\mathbb{P}} : f \mapsto \int f(x)k(x, \cdot)d\mathbb{P}(x). \tag{4.8.1}$$

The assumption that k is square integrable implies that $\mathbf{T}_{k,\mathbb{P}}$ is a bounded linear operator and $\|\mathbf{T}_{k,\mathbb{P}}\|_{\text{op}} \leq \|k\|_{\mathbb{P}}$. Moreover, this ensures that $\mathbf{T}_{k,\mathbb{P}}$ is compact (e.g., [36, Prop. 4.7]). If the kernel function is symmetric, then $\mathbf{T}_{k,\mathbb{P}}$ is a self-adjoint operator and, by the spectral theory for self-adjoint compact operators, $\mathbf{T}_{k,\mathbb{P}}$ has a countable sequence of eigenvalues $\lambda_1, \lambda_2, \ldots$ and orthonormal eigenfunctions $r_1, r_2, \ldots$ and can be represented by the series

$$\mathbf{T}_{k,\mathbb{P}}h = \sum_{i=1}^{\infty} \lambda_i \langle r_i, h \rangle r_i,$$

which converges in $L^2(\mathbb{P})$-norm. As a corollary of this result, any symmetric and square summable kernel function can be represented, in the sense of $L^2(\mathbb{P})$ convergence, by the series

$$k(x, y) = \sum_{i=1}^{\infty} \lambda_i r_i(x) r_i(y), \tag{4.8.2}$$

for x and y in the support of $\bar{\mathbb{P}}$. (For instance, see Riesz and Nagy [95, Sec. 97]). We say a kernel function is positive semidefinite if for any finite set of elements $x_1, \ldots, x_n$ the kernel matrix with entries $A_{ij} = k(x_i, x_j)$ is positive semidefinite. When the kernel function is symmetric, continuous, and positive semidefinite, then we have a slightly stronger statement about the representation of k by the eigenfunctions of $\mathbf{T}_{k,\mathbb{P}}$.

Theorem 4.8.1 (Mercer)**.** *Suppose that $\mathcal{X}$ is compact, and that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous, symmetric and positive semidefinite, then the representation* (4.8.2) *is uniformly convergent.*

For a proof in the simple case where $\mathcal{X}$ is an interval $[a, b] \subset \mathbb{R}$, see Riesz and Nagy [95, Sec. 98]. For a full, updated treatment see the book by Steinwart and Christmann [106].

Reproducing kernel Hilbert spaces

A *reproducing kernel Hilbert space* is a Hilbert space of real-valued functions on $\mathcal{X}$ with the property that for each $x \in \mathcal{X}$, the evaluation functional $f \mapsto f(x)$ is bounded. For any RKHS, there exists a unique positive semidefinite kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for any $h \in \mathcal{H}$ and $x \in \mathcal{X}$, we have $\langle h, k(x, \cdot)\rangle_{\mathcal{H}} = h(x)$. Conversely, given any positive semidefinite kernel function k, we can construct a RKHS, $\mathcal{H}$, in which $k_x := k(x, \cdot)$ acts as the representer of evaluation at x. We construct this Hilbert space as (the completion of) the set of finite linear combinations $\sum_{i=1}^{n} a_i k(x_i, \cdot)$ equipped with inner product defined by

$$\big\langle k_x, k_y \big\rangle = k(x, y). \tag{4.8.3}$$

One striking fact is that for any given any kernel k and distribution $\mathbb{P}$ on $\mathcal{X}$, the RKHS is isomorphic to an ellipsoid in $L^2(\mathbb{P})$. Consider the kernel integral operator $\mathbf{T}_{k,\mathbb{P}} : L^2(\mathbb{P}) \to L^2(\mathbb{P})$ defined in equation (4.8.1). Denote its eigenfunctions by $\{r_i\}_{i=1}^{\infty}$. Let $\mathbf{H}_{\mathbb{P}} : \mathcal{H} \to \mathcal{H}$ denote the integral operator on $\mathcal{H}$ defined by $\mathcal{H} \ni h \mapsto \int h(x) k_x d\mathbb{P}(x)$. If r_i satisfies $\|r_i\|_{\bar{\mathbb{P}}} = 1$ and $\mathbf{T}_{k,\mathbb{P}} r_i = \lambda_i r_i$, then $g_i := \sqrt{\lambda_i} r_i$ has unit norm in $\mathcal{H}$, and is an eigenfunction of $\mathbf{H}_{\mathbb{P}}$ with eigenvalue λ_i. By Mercer's theorem, for any probability measure $\mathbb{P}$ on $\mathcal{X}$, an element of the RKHS can be represented in terms of the eigenfunctions of $\mathbf{T}_{k,\mathbb{P}}$ by

$$h(x) = \Big\langle \sum_{i=1}^{\infty} \lambda_i r_i(x) r_i, h \Big\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \lambda_i r_i(x) \big\langle r_i, h \big\rangle_{\mathcal{H}}.$$

Hence the $\mathcal{H}$ norm of such an h is $\sum_{m=1}^{\infty} \lambda_i \langle r_i, h\rangle_{\mathcal{H}}^2$, showing that h is a member of $L^2(\mathbb{P})$. Consequently, the RKHS is isomorphic to an ellipsoid in $L^2(\mathbb{P})$. For more background on reproducing kernel Hilbert spaces, we refer the reader to various standard references [4, 99, 100, 106, 118].

Operator perturbation theorem

In this section, we state an operator perturbation theorem due to Stewart [107]. Let T be a self-adjoint Hilbert–Schmidt operator on a Hilbert space $\mathcal{H}$. Let $\mathcal{U} \subset \mathcal{H}$ be a subspace, let $\mathcal{U}^\perp$ denote its orthogonal complement, and let $\Pi_\mathcal{U}$ and $\Pi_{\mathcal{U}^\perp}$ denote the corresponding projection operators. For the decomposition of T according to $(\mathcal{U},\mathcal{U}^\perp)$, write

$$\begin{bmatrix} A & G^* \\ G & B \end{bmatrix}$$

where $A = \Pi_\mathcal{U} T \Pi_\mathcal{U}^*$, $G = \Pi_\mathcal{U} T \Pi_{\mathcal{U}^\perp}^*$, and $B = \Pi_{\mathcal{U}^\perp} T \Pi_{\mathcal{U}^\perp}^*$. Set

$$\alpha := \|G\|_{\text{HS}}, \qquad \beta := \inf\{|a-b| : a \in \sigma(A), b \in \sigma(B)\}.$$

Theorem 4.8.2 (Stewart)**.** *If $\frac{\alpha}{\beta} < \frac{1}{2}$, then there is an operator $E : \mathcal{U} \to \mathcal{U}^\perp$ such that* Range$(\Pi_\mathcal{U}^* + \Pi_{\mathcal{U}^\perp}^* E)$ *is an invariant subspace of T, and satisfying the bound $\|E\|_{\text{HS}} \leq 2\frac{\alpha}{\beta}$. Moreover, the spectrum of T is the disjoint union*

$$\sigma(T) = \sigma(A + G^*E) \cup \sigma(B - EG^*).$$

Cheeger's inequality

In this section, we state a version of Cheeger's inequality due to Lawler and Sokal [75]. Suppose that $\Pi : \mathcal{X} \times \Sigma \to [0,\infty)$ is a transition probability kernel on $(\mathcal{X},\Sigma)$ with invariant measure $\tilde{\mathbb{P}}_m$. The transition kernel Π induces a linear transformation on $L^2(\tilde{\mathbb{P}}_m)$ via $f \mapsto \int f(y)\Pi(\cdot, dy)$. Let $\lambda_2(\Pi)$ denote the second largest eigenvalue of this linear transformation. Define

$$\Gamma(\Pi) := \inf_{0<\tilde{\mathbb{P}}_m(S)<1} \frac{\int_S \Pi(x, S^c)\tilde{\mathbb{P}}_m(dx)}{\tilde{\mathbb{P}}_m(S)\tilde{\mathbb{P}}_m(S^c)}, \tag{4.8.4}$$

where the infimum is taken over all measurable sets S.

Theorem 4.8.3 (Cheeger's inequality)**.** *We have the following inequalities*

$$1 - \frac{\Gamma(\Pi)^2}{8} \geq \lambda_2(\Pi) \geq 1 - \Gamma(\Pi). \tag{4.8.5}$$

4.9 List of symbols

Symbol	Definition
φ	difficulty of clustering problem
$\mathcal{S}_{\text{max}}(\bar{\mathbb{P}})$	expected intercluster similarity
$\mathcal{C}(\bar{\mathbb{P}})$	coupling parameter
Γ	indivisibility of mixture components
$\bar{\mathbb{P}}$	mixture distribution on $\mathcal{X}$
$\bar{q}$	square root kernelized density for $\bar{\mathbb{P}}$
$\mathbb{P}_m$	the m-th mixture component on $\mathcal{X}$
w_m	the m-th mixture weight
q_m	square root kernelized density for $\mathbb{P}_m$
k	kernel function on $\mathcal{X} \times \mathcal{X}$
$\bar{k}$	$\bar{q}$-normalized kernel function on $\mathcal{X} \times \mathcal{X}$
k_m	q_m-normalized kernel function on $\mathcal{X} \times \mathcal{X}$
$\bar{\mathbb{P}}_n$	empirical mixture distribution on $\mathcal{X}$
$\Vert \cdot \Vert_{\text{op}}$	operator norm
$\Vert \cdot \Vert_{\text{HS}}$	Hilbert–Schmidt norm
$\Vert\cdot\Vert_{\mathbb{P}}$	$L^2(\mathbb{P})$ norm
$\langle \cdot, \cdot \rangle_{\mathbb{P}}$	$L^2(\mathbb{P})$ inner product
ρ	distance between subspaces
$\mathcal{Q}$	span of square root kernelized densities
$\mathcal{R}$	M-dimensional principal eigenspace of $\bar{\mathbf{T}}$
$\mathcal{G}$	M-dimensional principal eigenspace of $\bar{\mathbf{H}}$
$\tilde{\mathcal{G}}$	M-dimensional principal eigenspace of $\tilde{\mathbf{H}}$
$\tilde{\mathcal{V}}$	M-dimensional principal eigenspace of $\tilde{\mathbf{T}}$
$\mathcal{V}$	M-dimensional principal eigenspace of $\hat{\mathbf{T}}$
$\mathbf{T}_{k,\mathbb{P}}$	generic kernel operator with kernel k integrating against $\mathbb{P}$
$\bar{\mathbf{T}}$	Laplacian operator on $L^2(\bar{\mathbb{P}})$
$\tilde{\mathbf{T}}$	intermediate operator on $L^2(\bar{\mathbb{P}}_n)$
$\hat{\mathbf{T}}$	Laplacian matrix on $L^2(\bar{\mathbb{P}}_n)$
$\bar{\mathcal{H}}$	reproducing kernel Hilbert space for $\bar{k}$
$\tilde{\mathbf{H}}$	intermediate operator on $\bar{\mathcal{H}}$
$\bar{\mathbf{H}}$	Laplacian operator on $\bar{\mathcal{H}}$

Bibliography

[1] A. Agarwal et al. "Learning sparsely used overcomplete dictionaries via alternating minimization". In: *arXiv preprint arXiv:1310.7991* (2013).

[2] A. Amini et al. "Pseudo-likelihood methods for community detection in large sparse networks". In: *Annals of Statistics* 41.4 (2013), pp. 2097–2122.

[3] T. Ando. "Totally positive matrices". In: *Linear algebra and its applications* 90 (1987), pp. 165–219.

[4] N. Aronszajn. "Theory of Reproducing Kernels". In: *Transactions of the American Mathematical Society* 68.3 (1950), pp. 337–404.

[5] S. Arora et al. "Simple, efficient, and neural algorithms for sparse coding". In: *arXiv preprint arXiv:1503.00778* (2015).

[6] D. Arthur and S. Vassilvitskii. "k-means++: The advantages of careful seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms.* 2007, pp. 1027–1035.

[7] F. Bach. "Convex relaxations of structured matrix factorizations". In: *Technical report* (2013). eprint: arXiv:1309.3117.

[8] F. Bach, S. Lacoste-Julien, and G. Obozinski. "On the equivalence between herding and conditional gradient algorithms". In: *arXiv preprint arXiv:1203.4523* (2012).

[9] W.U. Bajwa et al. "Compressed channel sensing: A new approach to estimating sparse multipath channels". In: *Proc. IEEE* 98.6 (2010), pp. 1058–1076.

[10] R. Baraniuk. "Compressive Sensing [lecture notes]". In: *IEEE Signal Process Mag* 24.4 (2007), pp. 118–121.

[11] R. Baraniuk and P. Steeghs. "Compressive radar imaging". In: *In IEEE Radar Conf., Waltham, MA* (2007), pp. 128–133.

[12] P. L. Bartlett and S. Mendelson. "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results". In: *Journal of Machine Learning Research* 3 (Nov. 2002), pp. 463–482.

[13] P.L. Bartlett, O. Bousquet, and S. Mendelson. "Local Rademacher complexities". In: *Annals of Statistics* 33.4 (2005), pp. 1497–1537.

[14] D. Batenkov and Y. Yomdin. "Algebraic fourier reconstruction of piecewise smooth functions". In: *Math. Comput.* 81 (2012).

[15] M. Belkin and P. Niyogi. “Laplacian Eigenmaps for Dimensionality Reduction and Data Representation”. In: *Neural Computation* 15 (2003), pp. 1373–1396.

[16] T. Bendory. “Robust Recovery of Positive Stream of Pulses”. In: *http://arxiv.org/abs/1503.08782* (2015).

[17] T. Bendory, S. Dekel, and A. Feuer. “Robust Recovery of Stream of Pulses using Convex Optimization”. In: *http://arxiv.org/abs/1412.3262* (2014).

[18] E. Van den Berg and M. Friedlander. “Sparse optimization with least-squares constraints”. In: *SIAM Journal on Optimization* 21.4 (2011), pp. 1201–1229.

[19] B.N. Bhaskar, G. Tang, and B. Recht. “Atomic Norm Denoising with Applications to Line Spectral Estimation”. In: *IEEE Transactions on Signal Processing* 61.23 (2013), pp. 5987–5999.

[20] Biomedical Imaging Group. *Benchmarking of Single-Molecule Localization Microscopy Software*. 2013. URL: http://bigwww.epfl.ch/palm/.

[21] J.S. Bonifacino et al. “Imaging intracellular fluorescent proteins at nanometer resolution”. In: *Science* 313 (2006), pp. 1642–1645.

[22] R. Von Borries, C.J. Miosso, and C. Potes. “Compressed sensing using prior information”. In: *Computational Advances in Multi-Sensor Adaptive Processing, 2007. CAMPSAP 2007. 2nd IEEE International Workshop on.* IEEE. 2007, pp. 121–124.

[23] N. Boyd, G. Schiebinger, and B. Recht. “The alternating descent conditional gradient method for sparse inverse problems”. In: *Preprint.* (2015).

[24] K. Bredies and H.K. Pikkarainen. “Inverse problems in spaces of measures”. In: *ESAIM: Control, Optimisation and Calculus of Variations* 19 (01 Jan. 2013), pp. 190–218. ISSN: 1262-3377. DOI: 10.1051/cocv/2011205. URL: http://www.esaim-cocv.org/article_S1292811911002053.

[25] S. Burer and R. Monteiro. “A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization”. English. In: *Mathematical Programming* 95.2 (2003), pp. 329–357. ISSN: 0025-5610.

[26] E.J. Candès and C. Fernandez-Granda. “Super-resolution from noisy data”. In: *Journal of Fourier Analysis and Applications* 19.6 (2013), pp. 1229–1254.

[27] E.J. Candès and C. Fernandez-Granda. “Towards a mathematical theory of super resolution”. In: *Comm. Pure Appl. Math* (2013).

[28] E.J. Candès and B. Recht. “Exact matrix completion via convex optimization”. In: *Communications of the ACM* 55.6 (2012), pp. 111–119.

[29] E.J. Candès, J. Romberg, and T. Tao. “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information”. In: *IEEE Trans. Inf. Thy.* 52.2 (2006), pp. 489–509.

[30] E.J. Candès and M. Wakin. “An introduction to compressive sampling”. In: *IEEE Signal Process. Mag.* 25.2 (2008), pp. 21–30.

[31] Y. Cao and D. Chen. “Consistency of regularized spectral clustering”. In: *Applied and Computational Harmonic Analysis* 30.3 (2011), pp. 319–336.

[32] C. Carathéodory. “Ueber den Variabilitaetsbereich der Fourier’schen Konstanten von positiven harmonischen Funktionen”. In: *Rend. Circ. Mat.* 32 (1911), pp. 193–217.

[33] C. Carathéodory. “Ueber den Variabilitaetsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehman.” In: *Math. Ann.* 64 (1907), pp. 95–115.

[34] Y. de Castro and F. Gamboa. “Exact Reconstruction using Beurling Minimal Extrapolation”. In: *http://arxiv.org/abs/1103.4951* (2011).

[35] V. Chandrasekaran et al. “The Convex Geometry of Linear Inverse Problems.” In: *Foundations of Computational Mathematics* 12.6 (2012), pp. 805–849.

[36] J. Conway. *A Course in Functional Analysis*. 2nd edition. Graduate Texts in Mathematics (Book 96). Springer, 1997.

[37] W.E. Donath and A.J. Hoffman. “Lower Bounds for the Partitioning of Graphs”. In: *IBM J. Res. Develop.* 17.5 (1973), pp. 420–425.

[38] D. Donoho. “Compressed sensing”. In: *IEEE Trans. Inf. Thy.* 52.4 (2006), pp. 1289–1306.

[39] D. Donoho. “Superresolution via Sparsity Constraints”. In: *SIAM J. Math. Anal.* (1992).

[40] D. Donoho and P. Stark. “Uncertainty principles and signal recovery”. In: *SIAM J. Appl. Math* 49 (1989), pp. 906–931.

[41] P. Dragotti, M. Vetterli, and T. Blu. “Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets strang-fix”. In: *IEEE Transactions on Signal Processing* 55 (2007), pp. 1741–1757.

[42] M. Duarte and R. Baraniuk. “Spectral compressive sensing”. In: *Applied and Computational Harmonic Analysis* 35.1 (2013), pp. 111–129.

[43] V. Duval and G. Peyré. “Exact Support Recovery for Sparse Spikes Deconvolution”. In: *Foundations of Computational Mathematics* 15.5 (2015).

[44] K.S. Eckhoff. “Accurate reconstructions of functions of finite regularity from truncated fourier series expansions”. In: *Math. Comput* 64 (1995), pp. 671–690.

[45] C. Ekanadham, D. Tranchina, and E.P. Simoncelli. “Neural spike identifcation with continuous basis pursuit”. In: *Computational and Systems Neuroscience (CoSyNe), Salt Lake City, Utah* (2011).

[46] D. Evanko. “Primer: fluorescence imaging under the diffraction limit”. In: *Nature Methods* 6 (2009), pp. 19–20.

[47] A.C. Fannjiang, T. Strohmer, and P. Yan. “Compressed remote sensing of sparse objects”. In: *SIAM J. Imag. Sci.* 3.3 (2010), pp. 595–618.

[48] C. Fernandez-Granda. “Support detection in super-resolution”. In: *arXiv:1302.3921* (2013).

[49] M. Fiedler. “Algebraic connectivity of graphs”. In: *Czech. Math. J.* 23 (1973), pp. 298–305.

[50] M.P. Friedlander et al. “Recovering compressively sampled signals using partial support information”. In: *Information Theory, IEEE Transactions on* 58.2 (2012), pp. 1122–1134.

[51] J.-J. Fuchs. “Sparsity and uniqueness for some specific under-determined linear systems”. In: *In Acoustics, Speech, and Signal Processing* (2005).

[52] F.P. Gantmacher and M.G. Krein. “Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems”. In: *Revised English Ed. AMS Chelsea Pub. Providence, RI* (2002).

[53] A. Hansson, Z. Liu, and L. Vandenberghe. “Subspace System Identification via Weighted Nuclear Norm Optimization”. In: *CoRR* abs/1207.0023 (2012).

[54] Z. Harchaoui, A. Juditsky, and A. Nemirovski. “Conditional gradient algorithms for norm-regularized smooth convex optimization”. In: *Mathematical Programming* (2014), pp. 1–38.

[55] R. Heckel, V. Morgenshtern, and M. Soltanolkotabi. “Super-Resolution Radar”. In: *arXiv:1411.6272v2* (2015).

[56] M.A. Herman and T. Strohmer. “High-resolution radar via compressed sensing”. In: *IEEE Trans. Signal Process.* 57.6 (2009), pp. 2275–2284.

[57] S.T. Hess, T.P. Giriajan, and M.D. Mason. “Ultra-high resolution imaging by fluorescence pho- toactivation localization microscopy”. In: *Biophysical Journal* 91 (2006), pp. 4258–4272.

[58] R. Hettich and K. Kortanek. “Semi-infinite programming: theory, methods, and applications”. In: *SIAM review* 35.3 (1993), pp. 380–429.

[59] H. Hindi. “A tutorial on optimization methods for cancer radiation treatment planning”. In: *American Control Conference (ACC), 2013.* June 2013, pp. 6804–6816.

[60] M. Jaggi. “Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization”. In: *ICML* (2013.).

[61] M. Jaggi and M. Sulovsk. “A simple algorithm for nuclear norm regularized problems”. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10).* 2010, pp. 471–478.

[62] P. Jain, P. Netrapalli, and S. Sanghavi. “Low-rank matrix completion using alternating minimization”. In: *Proceedings of the forty-fifth annual ACM symposium on Theory of computing.* ACM. 2013, pp. 665–674.

[63] R. Jenssen. “Kernel Entropy Component Analysis”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (2010).

[64] S.G. Johnson. *The NLopt nonlinear-optimization package.* 2011. URL: http://ab-initio.mit.edu/nlopt.

[65] S. Karlin. “Total Positivity: Volume I”. In: *Stanford University Press* (1968).

[66] S. Karlin and W. Studden. “Tchebycheff Systems: with Applications in Analysis and Statistics”. In: *Wiley Interscience* (1967).

[67] R. Keshavan. "Efficient algorithms for collaborative filtering". PhD thesis. Stanford University, 2012.

[68] M.A. Khajehnejad et al. "Analyzing weighted minimization for sparse recovery with nonuniform sparse models". In: *Signal Processing, IEEE Transactions on* 59.5 (2011), pp. 1985–2001.

[69] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. ISBN: 0262013193, 9780262013192.

[70] V. Koltchinskii and E. Giné. "Empirical graph Laplacian approximation of Laplace–Beltrami operators: Large sample results". In: *IMS Lecture Notes–Monograph Series High Dimensional Probability* 51 (2006), pp. 238–259.

[71] M.G. Krein. "The ideas of P.L. Tchebycheff and A.A. Markov in the theory of limiting values of integrals and their futher development". In: *American Matematical Society Translations. Series 2.* 12 (1959).

[72] L. Zhu et al. "Faster storm using compressed sensing". In: *Nature Methods* 9 (2012), pp. 721–723.

[73] S. Lacoste-Julien, F. Lindsten, and F. Bach. "Sequential Kernel Herding: Frank-Wolfe Optimization for Particle Filtering". In: (2015).

[74] S. Laue. "A Hybrid Algorithm for Convex Semidefinite Optimization". In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. Ed. by John Langford and Joelle Pineau. ICML '12. Edinburgh, Scotland, GB: Omnipress, July 2012, pp. 177–184. ISBN: 978-1-4503-1285-1.

[75] G.F. Lawler and A.D. Sokal. "Bounds on the L^2 Spectrum for Markov Chains and Markov Processes: A Generalization of Cheeger's Inequality". In: *Transactions of the American Mathematical Society* 309 (1988), pp. 557–580.

[76] H.Y. Liu et al. "3D imaging in volumetric scattering media using phase-space measurements". In: *Opt. Express* 23.11 (June 2015), pp. 14461–14471.

[77] U. von Luxburg, M. Belkin, and O. Bousquet. "Consistency of Spectral Clustering". In: *Annals of Statistics* 36.2 (2008), pp. 555–586.

[78] D. Malioutov, M. Cetin, and A.S. Willsky. "A sparse signal reconstruction perspective for source localization with sensor arrays". In: *IEEE Trans. Signal Process* 53.8 (2005), pp. 3010–3022.

[79] M. Meila and J. Shi. "A Random Walks View of Spectral Segmentation". In: *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics* (2001).

[80] B.L.R. De Moor. *DaISy: Database for the Identification of Systems,*

[81] V.I. Morgenshtern and E.J. Candès. "Stable Super-Resolution of Positive Sources: the Discrete Setup". In: *arXiv:1504.00717* (2015).

[82] P. Netrapalli, P. Jain, and S. Sanghavi. "Phase retrieval using alternating minimization". In: *Advances in Neural Information Processing Systems*. 2013, pp. 2796–2804.

[83] A. Y. Ng, M.I. Jordan, and Y. Weiss. “On Spectral Clustering: Analysis and an algorithm”. In: *Advances in Neural Information Processing Systems (NIPS)* (2001).

[84] Nobelprize.org. *The Nobel Prize in Chemistry 2014*. URL: http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2014/.

[85] R. Ostrovsky et al. “The Effectiveness of *L*loyd-Type Methods for the k-Means Problem”. In: *Journal of the ACM* 59.6 (2012), 28:1–28:22.

[86] A. Pinkus. *Totally positive matrices*. Vol. 181. Cambridge University Press, 2010.

[87] B.G.R. de Prony. “Essai éxperimental et analytique: sur les lois de la dilatabilité de fluides élastique et sur celles de la force expansive de la vapeur de l’alkool, à différentes températures”. In: *Journal de l’École Polytechnique* 1.22 (1795), pp. 24–76.

[88] F. Pukelsheim. *Optimal design of experiments*. Vol. 50. siam, 1993.

[89] K.G. Puschmann and F. Kneer. “On super-resolution in astronomical imaging”. In: *Astronomy and Astrophysics* 436 (2005), pp. 373–378.

[90] N. Rao, P. Shah, and S. Wright. “Forward-Backward Greedy Algorithms for Atomic Norm Regularization”. In: *arXiv:1404.5692* (2014).

[91] H. Rauhut. “Random sampling of sparse trigonometric polynomials”. In: *Applied and Comput. Hamon. Anal.* 22.1 (2007), pp. 16–42.

[92] H. Rauhut and R. Ward. “Interpolation via weighted l_1 minimization”. In: *Applied and Computational Harmonic Analysis* (2015). To Appear. Preprint available at arxiv:1308.0759.

[93] B. Recht and C. Ré. “Parallel stochastic gradient algorithms for large-scale matrix completion”. In: *Mathematical Programming Computation* 5.2 (2013), pp. 201–226.

[94] M. Reed and B. Simon. *Methods of Modern Mathematical Physics*. Vol. I: Functional Analysis. Elsevier, 1980.

[95] F. Riesz and B. Sz.-Nagy. *Functional Analysis*. Dover, 1955.

[96] L. Rosasco, M. Belkin, and E. De Vito. “On Learning with Integral Operators”. In: *Journal of Machine Learning Research* 11 (2010), pp. 905–934.

[97] M.J. Rust, M. Bates, and X. Zhuang. “Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm)”. In: *Nature Methods* 3 (2006), pp. 793–796.

[98] D. Sage et al. “Quantitative evaluation of software packages for single-molecule localization microscopy”. In: *Nat Meth* advance online publication (June 15, 2015), pages. URL: http://dx.doi.org/10.1038/nmeth.3442.

[99] S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific and Technical, Harlow, UK, 1988.

[100] B. Scholkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[101] P. Shah et al. “Linear System Identification via Atomic Norm Regularization”. In: *arXiv:1204.0590* (2012).

[102] A. Shapiro. "Semi-infinite programming, duality, discretization and optimality conditions†". In: *Optimization* 58.2 (2009), pp. 133–161.

[103] J. Shi and J. Malik. "Normalized Cuts and Image Segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8 (2000), pp. 888–905.

[104] T. Shi, M. Belkin, and B. Yu. "Data spectroscopy: Eigenspaces of convolution operators and clustering". In: *Annals of Statistics* 37.6B (2009), pp. 3960–3984.

[105] J. Skaf and S. Boyd. "Techniques for exploring the suboptimal set". In: *Optimization and Engineering* 11.2 (2010), pp. 319–337.

[106] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

[107] G. W. Stewart. "Error Bounds for Approximate Invariant Subspaces, of Closed Linear Operator". In: *SIAM Journal on Numerical Analysis* 8.4 (Dec. 1971), pp. 796–808.

[108] P. Stoica and P. Babu. "Spice and likes: Two hyperparameter-free methods for sparse-parameter estimation". In: *Signal Processing* (2011).

[109] P. Stoica, P. Babu, and J. Li. "New method of sparse parameter estimation in separable models and its use for spectral analysis of irregularly sampled data". In: *IEEE Transactions on Signal Processing* 59.1 (2011), pp. 35–47.

[110] V. Tan and V. Goyal. "Estimating Signals With Finite Rate of Innovation From Noisy Samples: A Stochastic Algorithm". In: *IEEE Transactions on Signal Processing* 56.10 (Oct. 2008).

[111] G. Tang, B. Bhaskar, and B. Recht. "Near Minimax Line Spectral Estimation". In: *IEEE Transactions on Information Theory* (2014). To appear.

[112] G. Tang, B. Bhaskar, and B. Recht. "Sparse recovery over continuous dictionaries: Just discretize". In: *Asilomar* (2013).

[113] G. Tang et al. "Compressed Sensing off the Grid". In: *IEEE Transactions on Information Theory* 59.11 (2013), pp. 7465–7490.

[114] P.L. Tchebycheff. "On two theorems with respect to probabilities". In: *Zap. Akad. Nauk S.-Petersburg* 55 (1887), pp. 156–168.

[115] R.J. Tibshirani. "A General Framework for Fast Stagewise Algorithms". In: *arXiv preprint arXiv:1408.5801* (2014).

[116] N. Vaswani and W. Lu. "Modified-CS: Modifying compressive sensing for problems with partially known support". In: *Signal Processing, IEEE Transactions on* 58.9 (2010), pp. 4595–4607.

[117] M. Vetterli, P. Marziliano, and T. Blu. "Samplin signals with finite rate of innovation". In: *IEEE Transactions on Signal Processing* 50.6 (2002), pp. 1417–1428.

[118] G. Wahba. *Spline models for observational data.* CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, 1990.

[119] D. Yan, A. Chen, and M. I. Jordan. "Cluster Forests". In: *Computational Statistics and Data Analysis* 66 (2013), pp. 178–192.

[120] X. Zhang, Y. Yu, and D. Schuurmans. “Accelerated Training for Matrix-norm Regularization: A Boosting Approach”. In: *Advances in Neural Information Processing Systems 26 (NIPS)*. 2012.