Fitting Convex Sets to Data: Algorithms and Applications

Thesis by
Yong Sheng Soh

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2019
Defended August 29, 2018

Yong Sheng Soh
ORCID: 0000-0003-3367-1401

ACKNOWLEDGEMENTS

My first and foremost thanks go to my advisor Venkat Chandrasekaran. Reflecting back on my time in graduate school, Venkat has been the mentor that I needed, and perhaps at times, more so than the mentor that I necessarily wanted. Thank you for your dedication and hard work at training me to be an independent researcher. Thank you for your faith in me, and for pushing me hard to realize my potential. It was a tremendous privilege to work with you.

I am grateful to committee members Mathieu Desbrun, Babak Hassibi, Andrew Stuart, and Joel Tropp. Outside this thesis, they have been helpful informal mentors and teachers who have changed the way I think and do research. I also wish to thank Adam Wierman for supporting me throughout my time in graduate school, and for fostering an environment at CMS where I could thrive.

A big thank you to my friends at CMS: Ania Baetica, Max Budninsky, Desmond Cai, Utkan Candogan, Thomas Catanach, Niangjun Chen, Lingwen Gan, Ramya Korlakai, Yoke Peng Leong, Riley Murray, Yorie Nakahira, John Pang, and Armeen Taeb. It has been a blessing to learn from you, and be friends with you.

I am indebted to my family (my parents, my brothers, and my in-laws). No matter what the circumstances are, I am always able to count on their unconditional love and their unwavering support. Finally, I wish to thank my wife Li Ling Quek. I am deeply grateful for your sacrifice when you uprooted yourself so that we could be in the same place. The past few years we spent together exploring this country while going through rough patches were true blessings. Those years were like an unending honeymoon. I look forward to continuing our journey the rest of our lives.

ABSTRACT

This thesis concerns the geometric problem of finding a convex set that best fits a given dataset. Our question serves as an abstraction for data-analytical tasks arising in a range of scientific and engineering applications. We focus on two specific instances:

1. A key challenge that arises in solving inverse problems is ill-posedness due to a lack of measurements. A prominent family of methods for addressing such issues is based on augmenting optimization-based approaches with a convex penalty function so as to induce a desired structure in the solution. These functions are typically chosen using prior knowledge about the data. In Chapter 2, we study the problem of learning convex penalty functions directly from data for settings in which we lack the domain expertise to choose a penalty function. Our solution relies on suitably transforming the problem of learning a penalty function into a fitting task.

2. In Chapter 3, we study the problem of fitting tractably-described convex sets given the optimal value of linear functionals evaluated in different directions.

Our computational procedures for fitting convex sets are based on a broader framework in which we search among families of sets that are parameterized as linear projections of a fixed structured convex set. The utility of such a framework is that our procedures reduce to the computation of simple primitives at each iteration, and these primitives can be further performed in parallel. In addition, by choosing structured sets that are non-polyhedral, our framework provides a principled way to search over expressive collections of non-polyhedral descriptions; in particular, convex sets that can be described via semidefinite programming provide a rich source of non-polyhedral sets, and such sets feature prominently in this thesis.

We provide performance guarantees for our procedures. Our analyses rely on understanding geometrical aspects of determinantal varieties, building on ideas from empirical processes as well as random matrix theory. We demonstrate the utility of our framework with numerical experiments on synthetic data as well as applications in image denoising and computational geometry.

As secondary contributions, we consider the following:

1. In Chapter 4, we consider the problem of optimally approximating a convex set as a spectrahedron of a given size. Spectrahedra are sets that can be expressed as feasible regions of a semidefinite program.

2. In Chapter 5, we consider change-point estimation in a sequence of high-dimensional signals given noisy observations. Our method integrates classical approaches with a convex optimization-based step that is useful for exploiting structure in high-dimensional data.

PUBLISHED CONTENT AND CONTRIBUTIONS

Y. S. Soh and V. Chandrasekaran (2015). "High-Dimensional Change-Point Estimation: Combining Filtering with Convex Optimization". In *2015 IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, June 2015. DOI 10.1109/ISIT.2015.7282435
Y.S.S participated in the conception of the project, performed the analysis, undertook the numerical experiments, and participated in the writing of the manuscript.

Y. S. Soh and V. Chandrasekaran (2017). "High-Dimensional Change-Point Estimation: Combining Filtering with Convex Optimization". In *Applied and Computational Harmonic Analysis* 43,1, pp 122–147. DOI 10.1016/j.acha.2015.11.003
Y.S.S participated in the conception of the project, performed the analysis, undertook the numerical experiments, and participated in the writing of the manuscript.

Y. S. Soh and V. Chandrasekaran (2018). "Learning Semidefinite Regularizers". In *Foundations of Computational Mathematics*. DOI 10.1007/s10208-018-9386-z
Y.S.S participated in the conception of the project, performed the analysis, undertook the numerical experiments, and participated in the writing of the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

The heart of this thesis concerns the geometric problem of finding a convex set $C \subset \mathbb{R}^d$ that best fits a given dataset. Recall that a set C is *convex* if it satisfies the following property:

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in C \text{ for every } \mathbf{x}, \mathbf{y} \in C, \text{ and every } \lambda \in [0, 1].$$

The task of fitting a convex set to data serves as an abstraction of data-analytical tasks arising in a range of scientific and engineering applications. For example, in econometrics, one may wish to learn a convexity-based model to describe supply-demand levels of a specific commodity based on historical data; here, convexity-based considerations arise naturally as a result of fundamental principles such as marginal utility. In computational geometry, the computation of convex sets to encompass a collection of points in space is a routinely applied procedure, and it serves as a building block for describing more complicated objects.

The nature of the data that is presented here may take many forms; for instance, they may reveal direct information about C in the form of points lying on the boundary of C, or they may reveal indirect information about C in the form of optimal values of a collection of functions evaluated over C. The precise manner in which we fit a convex set depends on the type of data we receive as well as the purpose for which we fit these sets. In Chapters 2 and 3, we focus on two specific instances of fitting problems, and we summarize the contributions of these chapters as follows:

1. **Chapter 2: Learning Semidefinite Regularizers.** Regularization techniques are widely used to address ill-posedness in optimization-based approaches for solving inverse problems. These techniques frequently take the form of penalty functions that are augmented to the objective of our optimization instance. The purpose of introducing regularization is to induce a certain desired structure in the solutions to our inverse problems; for instance, the ℓ_1 norm is useful for inducing sparse structure in solutions, and the nuclear-norm is useful for promoting low-rank structure in matrices. In the most typical implementations of regularization techniques, the specific choice of regularizer (i.e. the

penalty function) is informed by domain-specific knowledge about the data; for instance, certain classes of signals are known to be well-approximated as being sparse in the Fourier domain, and a natural choice of regularizer that is effective at inducing such structure is the ℓ_1 norm with respect to the Fourier basis. Unfortunately, the challenge in many contemporary data-analytical tasks arising in scientific and engineering applications is that the data is frequently high-dimensional, and is presented in an unstructured manner. These challenges complicate the task of providing an informed choice of regularizer. To address these issues, we propose a framework for *learning* a suitable regularizer directly from data.

Our first contribution is to provide a conceptual link between the problem of learning a regularizer from data and the task of identifying a suitable atomic set. More precisely, atomic sets are collections of vectors – we refer to these as *atoms* – that specify a model for representing data succinctly. The relevance of atomic sets to our set-up is that they identify natural choices of regularizers that are effective at enforcing latent structure present in the data.

Our second contribution is to show that the simplest instantiation of learning a regularizer precisely corresponds to the more widely studied problem of 'dictionary learning' or 'sparse coding' – these concern the task of representing data as linear projections of sparse vectors. As we elaborate further in Chapter 2, the regularizers that we learn from dictionary learning correspond to identifying a finite collection of atoms for data, and are computable via linear programming.

Our third contribution is to propose and analyze an algorithm for learning regularizers which correspond to a specific infinite collection of atoms, and are computable via semidefinite programming. In particular, our framework naturally generalizes prior works in dictionary learning. A critical component of our procedure is that we apply an intermediate Operator Sinkhorn iterative step, and this is necessary to address identifiability issues that arise. The Operator Sinkhorn iterative step is the operator analog of a widely studied algorithm for matrix scaling known as Sinkhorn Scaling. Our main result in this chapter is a local linear convergence guarantee for our algorithm. The proof of our result relies on the analysis of affine rank minimization instances, the stability properties of Operator Sinkhorn and their relation to geometric aspects of determinantal varieties (in particular, the tangent spaces

with respect to these varieties), as well as ideas from random matrix theory.

Our fourth contribution is to demonstrate the utility of our framework for learning regularizers, and in particular, regularizers that correspond to an infinite collection of atoms. We consider a numerical task in which we denoise a collection of natural images corrupted by noise. Our results show that the denoisers obtained using our framework attain the same performance as denoisers obtained using prior works based on dictionary learning, but are computationally cheaper to evaluate.

2. **Chapter 3: Fitting Tractable Convex Sets to Support Function Evaluations.** In this chapter, we consider the problem of reconstructing a convex set given optimal values of linear functionals evaluated in different directions. More formally, given a vector $\mathbf{u} \in \mathbb{R}^d$ with unit Euclidean-norm, the *support function* of a convex set $C \subset \mathbb{R}^d$ evaluated in the direction $\mathbf{u} \in \mathbb{R}^d$ is defined as $h_C(\mathbf{u}) = \sup_{\mathbf{g} \in C} \langle \mathbf{g}, \mathbf{u} \rangle$. Our task is to estimate an unknown compact convex $C^\star \subset \mathbb{R}^d$ given access to its support function evaluations, which may further be corrupted by noise $\{(\mathbf{u}^{(i)}, y^{(i)}) : y^{(i)} = h_{C^\star}(\mathbf{u}^{(i)}) + \epsilon^{(i)}\}$.

   The task of reconstructing convex sets from their support function measurements arises in a range of applications; for instance, in tomographic applications, the extent of the absorption of parallel rays projected onto an object provides support information about the object. Previous approaches for estimating convex sets from support function measurements typically rely on estimators that minimize the error over all possible compact convex sets. Unfortunately, the drawback of such approaches is that they do not allow for the incorporation of prior structural information about the underlying set, and the resulting estimates become increasingly more complicated to describe as the number of measurements available grows. In addition, these estimates are frequently specified in terms of polyhedral descriptions, and are thus inadequate for expressing non-polyhedral sets.

   In this chapter, we describe a framework for estimating *tractable* convex sets from support function evaluations. Our approach is based on estimators that minimize the error over structured families of convex sets that are specified as linear images of concisely described sets in a higher-dimensional space that is not much larger than the ambient space. Examples of concisely described sets that we consider in this chapter include the simplex and the free spectrahedron. By parameterizing convex sets as linear images of concisely described

sets we achieve two outcomes: from a computational viewpoint, one can optimize linear functionals over estimators that we obtain as outputs from our procedures efficiently, and from an inferential perspective, we *regularize* for the complexity of the reconstruction.

We provide a geometric characterization of the asymptotic behavior of our estimators. Our analysis relies on a results that shows certain sets which admit semialgebraic descriptions are Vapnik-Chervonenkis (VC) classes. We apply our numerical procedures to a range of reconstruction tasks including computing a convex mesh of the human lung as well as (a variant of) fitting points on the unit-sphere in $\mathbb{R}^3$ so as to maximize separation. Our numerical experiments highlight the utility of our framework over previous approaches in settings in which the measurements available are noisy or small in number as well as those in which the underlying set to be reconstructed is non-polyhedral.

A Framework for Fitting. At the core of Chapters 2 and 3 is a broader framework for fitting convex sets based on searching over families specified as *linear projections* of 'structured' convex sets:

$$\mathscr{C} = \{A(C) : A \in L(\mathbb{R}^q, \mathbb{R}^d)\}. \tag{1.1}$$

For instance, in Chapter 3, we consider fitting with sets that are expressible as the projection of the simplex – this is the collection of non-negative vectors whose entries sum up to one:

$$\Delta^q := \{\mathbf{x} : \mathbf{x} = (x_1, \ldots, x_q)', \langle \mathbf{x}, \mathbf{1} \rangle = 1, x_i \geq 0, 1 \leq i \leq q\} \subset \mathbb{R}^q.$$

In addition, in Chapter 3, we also consider fitting with sets that are expressible as the projection of the free spectrahedron – this is the collection of positive semidefinite matrices whose eigenvalues sum up to one, and can be viewed as a suitable generalization of the simplex:

$$O^p := \{X : X = X', \mathrm{trace}(X) = \langle X, I \rangle = 1, \lambda_i(X) \geq 0, 1 \leq i \leq p\} \subset \mathbb{S}^p.$$

In Chapter 2, we are interested in fitting convex sets that arise as level sets of *norms*. Since norms are functions that are symmetric about the origin, it is natural to consider lifting sets C that are *centrally symmetric*, as the projections of such sets are also centrally symmetric. Here, recall that a set C is centrally symmetric if

$-\mathbf{x} \in C$ whenever $\mathbf{x} \in C$. As such, in Chapter 2, we primarily consider projections of the ℓ_1 ball (also known as the cross-polytope)

$$\left\{ \mathbf{x} : \mathbf{x} = (x_1, \ldots, x_q)', \sum_{i=1} |x_i| \leq 1 \right\} \subset \mathbb{R}^q,$$

as well as the nuclear-norm ball

$$\left\{ X : \sum_{i=1}^{p} \sigma_i(X) \leq 1 \right\} \subset \mathbb{R}^{p \times p}, \quad \sigma_i(\cdot) \text{ is the } i\text{-th singular value.}$$

The parameterization of families of convex sets as linear images of structured sets in (1.1) is a central idea in optimization as such families offer a powerful and expressive framework for describing feasible regions of convex programs. In the context of fitting convex sets to data, we search over families parameterized as (1.1) for similar reasons, and we elaborate on these as well as additional reasons in the following.

First, as noted above, the parameterization of convex sets as linear images of structured sets as in (1.1) offers a very expressive framework for describing convex sets. As a simple example to illustrate our point, every polytope with at most q vertices can be represented as the linear image of the simplex $\Delta^q \subset \mathbb{R}^q$ via an appropriate projection map $A \in L(\mathbb{R}^q, \mathbb{R}^d)$.

Second, by choosing structured convex sets that are non-polyhedral, our framework offers a principled approach for searching over collections of non-polyhedral sets. The capacity to accommodate non-polyhedral descriptions within our framework is a significant contribution of this thesis over prior works. As we elaborate in subsequent chapters, these prior works can be viewed suitably as instances of fitting convex sets with polyhedral descriptions to data.

Third, parameterizations of the form (1.1) provide a concrete approach to quantify the complexity of a set description, namely, in terms of the dimension of the lifting set C. We refer to such a quantity as the *lifting dimension*. The notion of using the lifting dimension as a measure of complexity of a set description is a core idea in the literature of 'lift-and-project' representations (also known as extended formulations) whereby one seeks compact descriptions of complicated objects by lifting to dimensions that are slightly higher than the ambient space. Fitting with sets that admit simple descriptions is an essential consideration within our framework as there are numerous instances in which we require these sets for downstream processing (the setting we consider in Chapter 2 is one such example). In such

settings, it is crucial that we are able to perform computations such as optimization over our fitted sets efficiently.

Fourth, we have effective computational strategies for searching over convex sets specified in the form of (1.1). The computational task of searching over families described as (1.1) naturally leads to an optimization instance over the space of linear maps $L(\mathbb{R}^q, \mathbb{R}^d)$. As we elaborate further in Chapters 2 and 3, such a task may be subsequently reformulated in terms of the structured factorization of a data matrix. The utility of such a reformulation is that we have access to a vast literature of prior works on computing matrix factorizations, and this is helpful for developing computational strategies for our fitting task. The procedures we develop in Chapters 2 and 3 are based on the idea that we compute structured factorizations by optimizing over one factor while keeping the other factor fixed. Such methods are frequently termed as 'alternating minimization'-based approaches, and they heavily rely on the minimization of each factor being relatively simple to compute. In the current context whereby we search over families parameterized by (1.1), the minimization of our associated matrix factorization instance over each factor requires computations that involve the facial structure of C. Consequently, it is advantageous to select choices of C for which its facial geometry is well-understood. Examples of sets that satisfy such a property include the simplex, the free spectrahedron, the ℓ_1 ball, and the nuclear norm ball.

Semidefinite Programming-Representable Convex Sets. A key emphasis of this thesis is that we fit using sets that can be described via semidefinite programming. Such sets are non-polyhedral, and they naturally generalize the collection of sets that can be described via linear programming, i.e., the collection of polyhedral sets.

A range of basic but important questions naturally arise in the context of fitting convex sets with semidefinite descriptions. For instance, one may wish to identify settings under which using semidefinite descriptions over polyhedral descriptions is advantageous. Another question is that one may wish to further understand the expressiveness of semidefinite descriptions. We investigate a simplified version of the latter question in Chapter 4, and we summarize the contributions of this chapter as follows:

1. **Chapter 4: Optimal Approximations of Convex Sets as Spectrahedra.** *Spectrahedra* are convex sets that can be specified via linear matrix inequalities. These are fundamental objects in the field of control theory as well

as optimization. The collection of spectrahedra is widely known to be very expressive; for instance, it includes the collection of all polyhedra. However, unlike the family of polyhedra, our understanding of spectrahedra is far more limited.

In this chapter, we focus on understanding the *expressiveness* of spectrahedra as convex sets. More precisely, we pose the following mathematical question: Given a compact convex $C \subset \mathbb{R}^d$, what is its optimal approximation as a spectrahedron of size k (i.e. the spectrahedron can be expressed in terms of a linear matrix inequality with matrices of size $k \times k$)? Building off ideas developed in Chapter 3, we develop numerical procedures for computing such approximations. Our computational tools are useful for understanding the expressiveness of spectrahedral sets from an approximation-theoretic viewpoint. We demonstrate numerical implementations of our procedure; in particular, we describe some examples of convex sets in which spectrahedra of small sizes offer a surprisingly high degree of approximation.

Secondary Contributions. As a secondary contribution, in Chapter 5, we consider the problem of estimating changes in a sequence of high-dimensional signals given noisy observations. We outline the contributions of Chapter 5 as follows:

1. **Chapter 5: High-Dimensional Change-Point Estimation: Combining Filtering with Convex Optimization.** Change-point estimation is the identification of abrupt changes in a sequence of observations. Such problems are routine in time-series analyses, and arise in a range of applications such as quality control and financial modeling. As in other inferential tasks encountered in contemporary settings, a key challenge underlying many modern change-point estimation problems is the increasingly large dimensionality of the underlying sequence of signals, and this leads both to computational difficulties as well as to complications with obtaining statistical consistency.

   A prominent family of methods for estimating the locations of change-points in a sequence of noisy *scalar*-valued observations is based on the filtered derivative. These methods begin with an averaging step applied over observations in a sliding window, followed by a computation of pairwise differences between successive averages, and finally the implementation of a thresholding step to estimate change-points. Unfortunately, the natural extension of the filtered derivative procedure to the *high-dimensional* setting leads to performance

guarantees for reliable change-point estimation that require the underlying signal to remain unchanged for long portions of the sequence.

In this chapter, we propose a new approach for estimating change-points in high-dimensional signals by integrating ideas from *atomic norm regularization* with the filtered derivative framework. The atomic norm regularization step is based on solving a convex optimization instance, and it exploits latent low-dimensional structure that is frequently found in signals encountered in practice. The specific choice of regularization may be selected based on prior knowledge about the data, or it may be learned from data using the ideas from Chapter 2. Our algorithm is well-suited to the high-dimensional setting both in terms of computational scalability and of statistical efficiency. More precisely, our main result shows that our method performs change-point estimation reliably as long as the product of the smallest-sized change (measured in terms of the Euclidean-norm-squared of the difference between signals at a change-point) and the smallest distance between change-points (as the number of time instances) are larger than a Gaussian width parameter that characterizes the low-dimensional complexity of the underlying signal sequence. Last, our method is applicable in online settings as it operates on small portions of the sequence of observations at a time.

1.1 Notation and Conventions

The styles of all variables and quantities follow these rules: a is a scalar, $\mathbf{a}$ is a vector, A is a matrix, and A is a linear operator mapping matrices to matrices. $\mathcal{A}$ is typically a set, while $\mathscr{A}$ is a *collection*. $\mathcal{A}$ denotes a linear map from the space of matrices to vectors.

We use d to refer to the *ambient* dimension, and we use q to refer to the *lifted* dimension. The exception is when we lift to the space of matrices, in which case we denote using $p \times p$ dimensions. We typically use s to denote the number of nonzero entries of a vector, i.e., *sparsity*, and r to denote *rank* of a matrix.

Last, $\|\cdot\|_{\ell_p}$ denotes the ℓ_p-norm; in particular, $\|\cdot\|_{\ell_2}$ denotes the Euclidean norm.

Chapter 2

LEARNING SEMIDEFINITE REGULARIZERS

2.1 Introduction

Regularization techniques are widely employed in the solution of inverse problems in data analysis and scientific computing due to their effectiveness in addressing difficulties due to ill-posedness. In their most common manifestation, these methods take the form of penalty functions added to the objective in optimization-based approaches for solving inverse problems. The purpose of the penalty function is to induce a desired structure in the solution, and these functions are specified based on prior domain-specific expertise. For example, regularization is useful for promoting smoothness, sparsity, low energy, and large entropy in solutions to inverse problems in image analysis, statistical model selection, and the geosciences [26, 30, 32, 35, 36, 48, 99, 116, 144]. In this paper, we study the question of *learning* suitable regularization functions from data in settings in which precise domain knowledge is not directly available. The regularizers obtained using our framework are specified as convex functions that can be computed efficiently via semidefinite programming, and therefore they can be employed in tractable convex optimization approaches for solving inverse problems.

We begin our discussion by highlighting the geometric aspects of regularizers that make them effective in promoting a desired structure. In particular, we focus on a family of convex regularizers that is useful for inducing a general form of sparsity in solutions to inverse problems. Sparse data descriptions provide a powerful formalism for specifying low-dimensional structure in high-dimensional data, and they feature prominently in a range of problem domains. For example, natural images are often well-approximated by a small number of wavelet coefficients, financial time series may be characterized by low-complexity factor models, and a small number of genetic markers may constitute a signature for disease. Concretely, suppose $\mathscr{A} \subset \mathbb{R}^d$ is a (possibly infinite) collection of elementary building blocks or atoms. Then $\mathbf{y} \in \mathbb{R}^d$ is said to have a sparse representation using the atomic set $\mathscr{A}$ if $\mathbf{y}$ can be expressed as follows:

$$\mathbf{y} = \sum_{i=1}^{k} c_i \mathbf{a}_i, \quad \mathbf{a}_i \in \mathscr{A}, c_i \geq 0,$$

for a relatively small number k. As an illustration, if $\mathscr{A} = \{\pm \mathbf{e}^{(j)}\}_{j=1}^{d} \subset \mathbb{R}^d$ is the collection of signed standard basis vectors in $\mathbb{R}^d$, then concisely described objects with these atoms are those vectors in $\mathbb{R}^d$ consisting of a small number of nonzero coordinates. Similarly, if $\mathscr{A}$ is the set of rank-one matrices, then the corresponding sparsely represented entities are low-rank matrices; see [35] for a more exhaustive collection of examples. An important virtue of sparse descriptions based on an atomic set $\mathscr{A}$ is that employing the *atomic norm* induced by $\mathscr{A}$ — the gauge function of the atomic set $\mathscr{A}$ — as a regularizer in inverse problems offers a natural convex optimization approach for obtaining solutions that have a sparse representation using $\mathscr{A}$ [35]. Continuing with the examples of vectors with few nonzero coordinates and of low-rank matrices, regularization with the ℓ_1 norm (the gauge function of the signed standard basis vectors) and with the matrix nuclear norm (the gauge function of the unit-Euclidean-norm rank-one matrices) are prominent techniques for promoting the corresponding sparse descriptions in solutions to inverse problems [30, 32, 36, 48, 53, 99, 116, 144]. The reason for the effectiveness of atomic norm regularization is the favorable facial structure of the convex hull of $\mathscr{A}$, which has the feature that all its low-dimensional faces contain points that have a sparse description using $\mathscr{A}$. Indeed, in many contemporary data analysis applications the solutions of regularized optimization problems with generic input data tend to lie on low-dimensional faces of sublevel sets of the regularizer [31, 48, 116]. Based on this insight, atomic norm regularization has been shown to be effective in a range of tasks such as statistical denoising, model selection, and system identification [20, 107, 127].

The difficulty with employing an atomic norm regularizer in practice is that one requires prior domain knowledge of the atomic set $\mathscr{A}$ – the extreme points of the atomic norm ball – that underlies a sparse description of the desired solution in an inverse problem. While such information may be available based on domain expertise in some problems (e.g., certain classes of signals having a sparse representation in a Fourier basis), identifying a suitable atomic set is challenging for many contemporary data sets that are high-dimensional and are typically presented to an analyst in an unstructured fashion. In this paper, we study the question of learning a suitable regularizer directly from observations $\{\mathbf{y}^{(j)}\}_{j=1}^{n} \subset \mathbb{R}^d$ of a collection of structured signals or models of interest. Specifically, as motivated by the preceding discussion, our objective is to identify a norm $\|\cdot\|$ in $\mathbb{R}^d$ such that each $\mathbf{y}^{(j)}/\|\mathbf{y}^{(j)}\|$ lies on a low-dimensional face of the unit ball of $\|\cdot\|$. An equivalent formulation of this question in terms of extreme points is that we want to obtain an atomic set $\mathscr{A}$ such

that each $\mathbf{y}^{(j)}$ has a sparse representation using $\mathscr{A}$; the corresponding regularizer is simply the atomic norm induced by $\mathscr{A}$. A norm with these characteristics is adapted to the structure contained in the data $\{\mathbf{y}^{(j)}\}_{j=1}^n$, and it can be used subsequently as a regularizer in inverse problems to promote solutions with the same type of structure as in the collection $\{\mathbf{y}^{(j)}\}_{j=1}^n$.

When considered in full generality, our question is somewhat ill-posed for several reasons. First, if $\|\cdot\|$ is a norm that satisfies the properties described above with respect to the data $\{\mathbf{y}^{(j)}\}_{j=1}^n$, then so does $\alpha\|\cdot\|$ for any positive scalar α. This issue is addressed by learning a norm from a suitably scaled class of regularizers. A second source of difficulty is that the Euclidean norm $\|\cdot\|_{\ell_2}$ trivially satisfies our requirements for a regularizer as each $\mathbf{y}^{(j)}/\|\mathbf{y}^{(j)}\|_{\ell_2}$ is an extreme point of the Euclidean norm ball in $\mathbb{R}^d$; indeed, this is the regularizer employed in ridge regression. The atomic set in this case is the collection of all points with Euclidean norm equal to one, i.e., the dimension of this set is $d-1$. However, data sets in many applications throughout science and engineering are well-approximated as sparse combinations of elements of atomic sets of much smaller dimension [12, 26, 35, 45, 83, 104, 111]. Identifying such lower-dimensional atomic sets is critical in inverse problems arising in high-dimensional data analysis in order to address the curse of dimensionality; in particular, as discussed in some of these preceding references, the benefits of atomic norm regularization in problems with large ambient dimension d are a consequence of measure concentration phenomena that crucially rely on the small dimensionality of the associated atomic set in comparison to d. We circumvent this second difficulty in learning a regularizer by considering atomic sets with appropriately bounded dimension. A third challenge with our question as it is stated is that the gauge function of the set $\{\pm\mathbf{y}^{(j)}/\|\mathbf{y}^{(j)}\|_{\ell_2}\}_{j=1}^n$ also satisfies the requirements for a suitable atomic norm as each $\mathbf{y}^{(j)}/\|\mathbf{y}^{(j)}\|_{\ell_2}$ is an extreme point of the unit ball of this regularizer. However, such a regularizer suffers from overfitting and does not generalize well as it is excessively tuned to the data set $\{\mathbf{y}^{(j)}\}_{j=1}^n$. Further, for large n this gauge function becomes intractable to characterize, and it does not offer a computationally efficient approach for regularization. We overcome this complication by considering regularizers that have effectively parametrized sets of extreme points, and consequently are tractable to compute.

The problem of learning a suitable polyhedral regularizer – an atomic norm with a unit ball that is a polytope – from data points $\{\mathbf{y}^{(j)}\}_{j=1}^n$ corresponds to identifying an appropriate *finite* atomic set to concisely describe each $\mathbf{y}^{(j)}$. This problem is

equivalent to the question of 'dictionary learning' (also called 'sparse coding') on which there is a substantial amount of prior work [2–4, 8, 9, 11, 72, 104, 122, 123, 134, 139, 140, 148] (see also the survey articles in [51, 96]). To see this connection, suppose without loss of generality that we parametrize a finite atomic set via a matrix $L \in \mathbb{R}^{d\times q}$ so that the columns of L and their negations specify the atoms. The associated atomic norm ball is the image under L of the ℓ_1 ball in $\mathbb{R}^q$. The columns of L are typically scaled to have unit Euclidean norm to address the scaling issues mentioned previously (see Section 2.2.4). The number of columns q controls the complexity of the atomic set as well as the computational tractability of describing the atomic norm, and is permitted to be larger than d (i.e., the 'overcomplete' regime). With this parametrization, learning a polyhedral regularizer to promote the type of structure contained in $\{\mathbf{y}^{(j)}\}_{j=1}^n$ may be viewed as obtaining a matrix L (given a target number of columns q) such that each $\mathbf{y}^{(j)}$ is well-approximated as $L\mathbf{x}^{(j)}$ for a vector $\mathbf{x}^{(j)} \in \mathbb{R}^q$ with few nonzero coordinates. Computing such a representation of the data is precisely the objective in dictionary learning, although this problem is typically not phrased as a quest for a polyhedral regularizer in the literature. We remark further on some recent algorithmic developments in dictionary learning in Sections 2.1.3.1 and 2.2.4, and we contrast these with the methods proposed in the present paper.

2.1.1 From Polyhedral to Semidefinite Regularizers

The objective of this paper is to investigate the problem of learning more general non-polyhedral atomic norm regularizers; in other words, the associated set of extreme points may be *infinite*. On the approximation-theoretic front, infinite atomic sets offer the possibility of concise descriptions of data sets with much richer types of structure than those with a sparse representation using finite atomic sets; in turn, the associated regularizers could promote a broader class of structured solutions to inverse problems than polyhedral regularizers. On the computational front, many families of convex optimization problems beyond linear programs can be solved tractably and reliably [103]. However, building on the challenges outlined previously, there are two important factors in identifying non-polyhedral regularizers from data. First, it is crucial that any infinite atomic set $\mathscr{A}$ we consider has an effective parametrization so that it is tractable to characterize data that have a sparse representation using the elements of $\mathscr{A}$. Second, we require that the convex hull of the atomic set $\mathscr{A}$ has an efficient description so that the associated atomic norm provides a computationally tractable regularizer. As described next, we address these

concerns by considering atomic sets that are efficiently parametrized as algebraic varieties (of a particular form), and that have convex hulls with tractable semidefinite descriptions. Thus, previous efforts in the dictionary learning literature on identifying finite atomic sets may be viewed as learning zero-dimensional ideals, whereas our approach corresponds to learning atomic sets that are larger-dimensional varieties. From a computational viewpoint, dictionary learning provides atomic norm regularizers that are computed via linear programming, while our framework leads to semidefinite programming regularizers. Consequently, although our framework is based on a much richer family of atomic sets in comparison with the finite sets considered in dictionary learning, we still retain efficiency of parametrization and computational tractability based on semidefinite representability.

Formally, we consider atomic sets in $\mathbb{R}^d$ that are images of rank-one matrices:

$$\mathscr{A}_p(\mathcal{L}) = \left\{\mathcal{L}(\mathbf{u}\mathbf{v}') \mid \mathbf{u}, \mathbf{v} \in \mathbb{R}^p,\ \|\mathbf{u}\|_{\ell_2} = 1, \|\mathbf{v}\|_{\ell_2} = 1\right\}, \tag{2.1}$$

where $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ specifies a linear map. We focus on settings in which the dimension p is such that $p^2 > d$, so the atomic sets $\mathscr{A}_p(\mathcal{L})$ that we study in this paper are projections of rank-one matrices from a larger-dimensional space (in analogy to the overcomplete regime in dictionary learning). By construction, elements of $\mathbb{R}^d$ that have a sparse representation using the atomic set $\mathscr{A}_p(\mathcal{L})$ are those that can be specified as the image under $\mathcal{L}$ of *low-rank matrices* in $\mathbb{R}^{p\times p}$. As the convex hull of unit-Euclidean-norm rank-one matrices in $\mathbb{R}^{p\times p}$ is the nuclear norm ball in $\mathbb{R}^{p\times p}$, the corresponding atomic norm ball is given by:

$$\operatorname{conv}\left(\mathscr{A}_p(\mathcal{L})\right) = \left\{\mathcal{L}(X) \mid X \in \mathbb{R}^{p\times p},\ \|X\|_\star \leq 1\right\}, \tag{2.2}$$

where $\|X\|_\star := \sum_i \sigma_i(X)$. As the nuclear norm ball has a tractable semidefinite description [53, 116], the atomic norm induced by $\mathscr{A}_p(\mathcal{L})$ can be computed efficiently using semidefinite programming.

Given a collection of data points $\{\mathbf{y}^{(j)}\}_{j=1}^n \subset \mathbb{R}^d$ and a target dimension p, our goal is to find a linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ such that each $\mathbf{y}^{(j)}$, upon normalization by the gauge function of $\mathscr{A}_p(\mathcal{L})$, lies on a low-dimensional face of $\operatorname{conv}(\mathscr{A}_p(\mathcal{L}))$. For each $\mathbf{y}^{(j)}$ to have this property, it must have a sparse representation using the atomic set $\mathscr{A}_p(\mathcal{L})$; that is, there must exist a low-rank matrix $X^{(j)} \in \mathbb{R}^{p\times p}$ with $\mathbf{y}^{(j)} = \mathcal{L}(X^{(j)})$. The matrix $X^{(j)}$ provides a concise description of $\mathbf{y}^{(j)} \in \mathbb{R}^d$ in the higher-dimensional space $\mathbb{R}^{p\times p}$. Consequently, the problem of learning a semidefinite-representable regularizer with a unit ball that is a linear image of the

	Dictionary learning	Our work
Atomic set	$\{\pm L\mathbf{e}^{(i)} \mid \mathbf{e}^{(i)} \in \mathbb{R}^q$ is the i'th standard basis vector$\}$ $L : \mathbb{R}^q \to \mathbb{R}^d$ (linear map)	$\{\mathcal{L}(\mathbf{u}\mathbf{v}') \mid \mathbf{u}, \mathbf{v} \in \mathbb{R}^p, \|\mathbf{u}\|_{\ell_2} = \|\mathbf{v}\|_{\ell_2} = 1\}$ $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ (linear map)
Algebraic/geometric structure of atoms	Zero-dimensional ideal	Image of determinantal variety
Concisely specified data using atomic set	Image under L of sparse vectors	Image under $\mathcal{L}$ of low-rank matrices
Atomic norm ball	$\left\{L\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^q,\ \|\mathbf{x}\|_{\ell_1} \le 1\right\}$	$\{\mathcal{L}(X) \mid X \in \mathbb{R}^{p\times p},\ \|X\|_\star \le 1\}$
Computing atomic norm regularizer	Linear programming	Semidefinite programming
Learning regularizer from data $\{\mathbf{y}^{(j)}\}_{j=1}^n$	Identify L and sparse $\mathbf{x}^{(j)} \in \mathbb{R}^q$ such that $\mathbf{y}^{(j)} \approx L\mathbf{x}^{(j)}$ for each j	Identify $\mathcal{L}$ and low-rank $X^{(j)} \in \mathbb{R}^{p\times p}$ such that $\mathbf{y}^{(j)} \approx \mathcal{L}(X^{(j)})$ for each j

Table 2.1: A comparison between prior work on dictionary learning and the present paper.

nuclear norm ball may be phrased as one of *matrix factorization*. In particular, let $Y = [\mathbf{y}^{(1)}|\cdots|\mathbf{y}^{(n)}] \in \mathbb{R}^{d\times n}$ denote the data matrix, and let $\mathcal{L}_i \in \mathbb{R}^{p\times p}$, $i = 1,\ldots,d$ be the matrix that specifies the linear functional corresponding to the i'th component of a linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$. Then our objective can be viewed as one of finding a collection of matrices $\{\mathcal{L}_i\}_{i=1}^d \subset \mathbb{R}^{p\times p}$ specifying linear functionals and a set of low-rank matrices $\{X^{(j)}\}_{j=1}^n \subset \mathbb{R}^{p\times p}$ specifying concise descriptions such that:

$$Y_{i,j} = \langle \mathcal{L}_i, X^{(j)} \rangle \quad i = 1,\ldots,d,\ j = 1,\ldots,n. \tag{2.3}$$

Here $\langle A, B\rangle = \mathrm{tr}(A'B)$ denotes the trace inner product between matrices. Note the distinction with dictionary learning in which one seeks a factorization of the data matrix Y such that the $X^{(j)}$'s are sparse vectors as opposed to low-rank matrices as in our approach. Figure 2.1 summarizes the key differences between dictionary learning and the present paper.

2.1.2 An Alternating Update Algorithm for Matrix Factorization

A challenge with identifying a semidefinite regularizer by factoring a given data matrix as in (2.3) is that such a factorization is not unique. Specifically, consider any linear map $\mathsf{M} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ that is a rank-preserver, i.e., $\mathrm{rank}(\mathsf{M}(X)) = \mathrm{rank}(X)$ for all $X \in \mathbb{R}^{p\times p}$; examples of rank-preservers include operators that act via conjugation by non-singular matrices and the transpose operation. If each $\mathbf{y}^{(j)} = \mathcal{L}(X^{(j)})$ for a linear map $\mathcal{L}$ and low-rank matrices $\{X^{(j)}\}_{j=1}^n$, then we also have

that each $\mathbf{y}^{(j)} = \mathcal{L} \circ \mathsf{M}^{-1}(\mathsf{M}(X^{(j)}))$, where by construction each $X^{(j)}$ has the same rank as the corresponding $\mathsf{M}(X^{(j)})$. This non-uniqueness presents a difficulty as the image of the nuclear norm ball under a linear map $\mathcal{L}$ is, in general, different than it is under $\mathcal{L} \circ \mathsf{M}^{-1}$ for an arbitrary rank-preserver M. Consequently, due to its invariances the factorization (2.3) does not uniquely specify a regularizer. We investigate this point in Section 2.2.2 by analyzing the structure of rank-preserving linear maps, and we describe an approach to associate a unique regularizer to a family of linear maps obtained from equivalent factorizations. Our method entails putting linear maps in an appropriate 'canonical' form using the Operator Sinkhorn iterative procedure, which was developed by Gurvits to solve certain quantum matching problems [75]; this algorithm is an operator analog of the diagonal congruence scaling technique for nonnegative matrices developed by Sinkhorn [132].

In Section 2.2 we describe an alternating update algorithm to compute a factorization of the form (2.3). With the $\mathcal{L}_i$'s fixed, updating the $X^{(j)}$'s entails the solution of affine rank minimization problems. Although this problem is intractable in general [102], in recent years several tractable heuristics have been developed and proven to succeed under suitable conditions [65, 82, 116]. With the $X^{(j)}$'s fixed, the $\mathcal{L}_i$'s are updated by solving a least-squares problem followed by an application of the Operator Sinkhorn iterative procedure to put the map $\mathcal{L}$ in a canonical form as described above. Our alternating update approach is a generalization of methods that are widely employed in dictionary learning for identifying finite atomic sets (see Section 2.2.4).

Section 2.3 contains the main theorem of this paper on the local linear convergence of our alternating update algorithm. Specifically, suppose a collection of data points $\{\mathbf{y}^{(j)}\}_{j=1}^{n} \subset \mathbb{R}^d$ is generated as $\mathbf{y}^{(j)} = \mathcal{L}^\star(X^{(j)\star})$, $j = 1, \ldots, n$ for a linear map $\mathcal{L}^\star : \mathbb{R}^{p \times p} \to \mathbb{R}^d$ that is nearly isometric restricted to low-rank matrices (formally, $\mathcal{L}^\star$ satisfies a *restricted isometry property* [116]) and a collection $\{X^{(j)\star}\}_{j=1}^{n} \subset \mathbb{R}^{p \times p}$ of low-rank matrices that is isotropic in a well-defined sense. Given the data $\{\mathbf{y}^{(j)}\}_{j=1}^{n}$ as input, our alternating update approach is locally linearly convergent to a linear map $\hat{\mathcal{L}} : \mathbb{R}^{p \times p} \to \mathbb{R}^d$ with the property that the image of the nuclear norm ball in $\mathbb{R}^{p \times p}$ under $\hat{\mathcal{L}}$ is equal to its image under $\mathcal{L}^\star$, i.e., our procedure identifies the appropriate regularizer that promotes the type of structure contained in the data $\{\mathbf{y}^{(j)}\}_{j=1}^{n}$; see Theorem 2.3.5. Our analysis relies on geometric aspects of determinantal varieties (in particular tangent spaces with respect to these varieties) and their relation to stability properties of Operator Sinkhorn scaling.

We demonstrate the utility of our framework with a series of experimental results on synthetic as well as real data in Section 2.4.

2.1.3 Related Work

2.1.3.1 Dictionary Learning

As outlined above, our approach for learning a regularizer from data may be viewed as a semidefinite programming generalization of dictionary learning. The alternating update algorithm we propose in Section 2.2.3 for computing a factorization (2.3) generalizes similar methods previously developed for dictionary learning [2, 4, 9, 104] (see Section 2.2.4), and the local convergence analysis of our algorithm in Section 2.3 also builds on previous analyses for dictionary learning [2, 9]. In contrast to these previous results, the development and the analysis of our method in the present paper are more challenging due to the invariances and associated identifiability issues underlying the factorization (2.3), which necessitate the incorporation of the Operator Sinkhorn scaling procedure in our algorithm.

An unresolved matter in our paper – one that has been investigated previously in the context of dictionary learning – is the question of a suitable initialization for our algorithm. In particular, our theory states that our algorithm exhibits linear convergence to the desired solution provided the initial guess is sufficiently close to a linear map that specifies the correct regularizer (in an appropriate metric). We employ random initializations in our experiments with real data in Section 2.4.2, and these are useful in identifying effective semidefinite regularizers that outperform polyhedral regularizers obtained via dictionary learning. Random initialization is the most common technique utilized in practice in dictionary learning as well as in many other structured matrix factorization problems arising in data analysis. To build support for this idea, several researchers have proven that random initialization succeeds with high probability in recovering a desired factorization under suitable conditions in a number of problems [64, 141], including in a restricted form of dictionary learning [139, 140] in which the polyhedral regularizer is specified as the image of the ℓ_1 ball under an invertible linear map (as described previously, dictionary learning in full generality allows for polyhedral regularizers that may be specified as an image of the ℓ_1 ball under a many-to-one linear map). In a different direction, some recent papers also describe data-driven initialization strategies for dictionary learning based on variants of clustering [3, 8]. It would be of interest to develop both these sets of ideas in our context, and we comment on this point in

Section 2.5.

2.1.3.2 Lifts of Convex Sets

A second body of work with which our paper is conceptually related is the literature on lift-and-project representations (or extended formulations) of convex sets. A tractable lift-and-project representation refers to a description of a 'complicated' convex set in $\mathbb{R}^d$ as the projection of a more concisely specified convex set in $\mathbb{R}^{d'}$, with the lifted dimension d' not being too much larger than the original dimension d. As discussed in [70, 152], obtaining a suitably structured factorization – of a different nature than that considered in the present paper – of the *slack matrix* of a polytope (and more generally, of the slack operator of a convex set) corresponds to identifying an efficient lift-and-project description of the polytope. On the other hand, we seek a structured factorization of a *data matrix* to identify a convex set (i.e., the unit ball of a regularizer) with an efficient extended formulation and with the additional requirement that the data points (upon suitable scaling) lie on low-dimensional faces of the set. This latter stipulation arises in our context from data analysis considerations, and it is a distinction between our setup and the optimization literature on extended formulations.

2.1.3.3 Sinkhorn Scaling

A third topic with which our paper has synergies – and to which we make contributions in the course of our analysis – is the literature on Sinkhorn scaling. This algorithm is an iterative procedure for transforming an entrywise nonnegative matrix to a doubly stochastic matrix by diagonal congruence scaling [132]. There is a substantial body of work on the properties of this algorithm (see [80] and the references therein) as well as on its applications in domains such as combinatorial optimization (approximating the permanent of a matrix [92]) and data analysis (efficiently computing distances between probability distributions [40]). The operator analog of Sinkhorn scaling was developed by Gurvits and this work was motivated by certain operator analogs of the bipartite matching problem that arise in matroid theory [75]. To the best of our knowledge, our work represents the first application of Operator Sinkhorn scaling in a problem in data analysis. Further, in our investigation of the properties of Algorithm 1, we describe results on the stability of Operator Sinkhorn scaling; these may be of independent interest beyond the specific context of our paper (see Appendix A.3).

2.1.4 Paper Outline

In Section 2.2 we discuss our alternating update algorithm for computing the factorization (2.3) based on an analysis of the invariances arising in (2.3). Section 2.3 gives the main theoretical result concerning the local linear convergence of the algorithm described in Section 2.2, and Section 2.4 describes numerical results obtained using our algorithm. We conclude with a discussion of further research directions in Section 2.5.

Notation We denote the Euclidean norm by $\|\cdot\|_{\ell_2}$. We denote the operator or spectral norm by $\|\cdot\|_2$. The k'th largest singular value of a linear map is denoted by $\sigma_k(\cdot)$, and the largest and smallest eigenvalues of a self-adjoint linear map are denoted by $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ respectively. The space of $p \times p$ symmetric matrices is denoted $\mathbb{S}^p$ and the set of $p \times p$ symmetric positive-definite matrices is denoted $\mathbb{S}^p_{++}$. The projection map onto a subspace $\mathcal{V}$ is denoted $\mathcal{P}_{\mathcal{V}}$. The restriction of a linear map M to a subspace $\mathcal{V}$ is denoted by $M_{\mathcal{V}}$. Given a self-adjoint linear map $M : \mathcal{V} \to \mathcal{V}$ with $\mathcal{V}$ being a subspace of a vector space $\bar{\mathcal{V}}$, we denote the extension of M to $\bar{\mathcal{V}}$ by $[M]_{\bar{\mathcal{V}}} : \bar{\mathcal{V}} \to \bar{\mathcal{V}}$; the component in $\mathcal{V}$ of the image of any $\mathbf{x} \in \bar{\mathcal{V}}$ under this map is $M\mathcal{P}_{\mathcal{V}}(\mathbf{x})$, while the component in $\mathcal{V}^\perp$ is the origin. Given a vector space $\mathcal{V}$, we denote the set of linear operators from $\mathcal{V}$ to $\mathcal{V}$ by $\mathrm{End}(\mathcal{V})$. Given matrices $A, B \in \mathbb{R}^{p\times p}$, the linear map $A \boxtimes B \in \mathrm{End}(\mathbb{R}^{p\times p})$ is specified as $A \boxtimes B : X \to \langle B, X\rangle A$. The Kronecker product between two linear maps is specified using the standard $\otimes$ notation. For a collection of matrices $\mathscr{X} := \{X^{(j)}\}_{j=1}^n \subset \mathbb{R}^{p\times p}$, the covariance is specified as $\mathrm{cov}(\mathscr{X}) = \frac{1}{n}\sum_{j=1}^n X^{(j)} \boxtimes X^{(j)}$. Two quantities associated to this covariance that play a role in our analysis are $\Lambda(\mathscr{X}) = \frac{1}{2}(\lambda_{\max}(\mathrm{cov}(\mathscr{X})) + \lambda_{\min}(\mathrm{cov}(\mathscr{X})))$ and $\Delta(\mathscr{X}) = \frac{1}{2}(\lambda_{\max}(\mathrm{cov}(\mathscr{X})) - \lambda_{\min}(\mathrm{cov}(\mathscr{X})))$. Given a matrix $X \in \mathbb{R}^{p\times p}$ of rank r, the tangent space at X with respect to the algebraic variety of $p \times p$ matrices of rank at most r is specified as[1]:

$$\mathcal{T}(X) = \{XA + BX \mid A, B \in \mathbb{R}^{p\times p}\}.$$

2.2 An Alternating Update Algorithm for Learning Semidefinite Regularizers

In this section we describe an alternating update algorithm to factor a given data matrix $Y = [\mathbf{y}^{(1)}|\cdots|\mathbf{y}^{(n)}] \in \mathbb{R}^{d\times n}$ as in (2.3). As discussed previously, the difficulty with obtaining a semidefinite regularizer using a factorization (2.3) is the existence of infinitely many equivalent factorizations due to the invariances underlying (2.3).

[1] A rank-r matrix $X \in \mathbb{R}^{p\times p}$ is a smooth point with respect to the variety of $p \times p$ matrices of rank at most r.

We begin by investigating and addressing this issue in Sections 2.2.1 and 2.2.2, and then we discuss our algorithm to obtain a regularizer in Section 2.2.3. We contrast our method with techniques that have previously been developed in the context of dictionary learning in Section 2.2.4.

2.2.1 Identifiability Issues

Building on the discussion in the introduction, for a linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ obtained from the factorization (2.3) and for any linear rank-preserver $\mathsf{M} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$, there exists an equivalent factorization in which the linear map is $\mathcal{L} \circ \mathsf{M}$ (note that M^{-1} is also a rank-preserver if M is a rank-preserver). As the image of the nuclear norm ball in $\mathbb{R}^{p\times p}$ is not invariant under an arbitrary rank-preserver, a regularizer cannot be obtained uniquely from a factorization due to the existence of equivalent factorizations that lead to non-equivalent regularizers. To address this difficulty, we describe an approach to associate a *unique* regularizer to a family of linear maps obtained from equivalent factorizations. We begin by analyzing the structure of rank-preserving linear maps based on the following result [98]:

Theorem 2.2.1 *([98, Theorem 1],[147, Theorem 9.6.2]) An invertible linear operator* $\mathsf{M} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ *is a rank-preserver if and only if* M *is of one of the following two forms for non-singular matrices* $W_1, W_2 \in \mathbb{R}^{p\times p}$*:* $\mathsf{M}(X) = W_1 X W_2$ *or* $\mathsf{M}(X) = W_1 X' W_2$.

This theorem brings the preceding discussion into sharper focus, namely, that the lack of identifiability boils down to the fact that the nuclear norm is not invariant under conjugation of its argument by arbitrary non-singular matrices. However, we note that the nuclear norm ball is invariant under the transpose operation and under conjugation by orthogonal matrices. This observation leads naturally to the idea of employing the *polar decomposition* to describe a rank-preserver:

Corollary 2.2.2 *Every rank-preserver* $\mathsf{M} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ *can be uniquely decomposed as* $\mathsf{M} = \mathsf{M}^{\mathrm{or}} \circ \mathsf{M}^{\mathrm{pd}}$ *for rank-preservers* $\mathsf{M}^{\mathrm{pd}} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ *and* $\mathsf{M}^{\mathrm{or}} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ *with the following properties:*

- *The operator* M^{pd} *is specified as* $\mathsf{M}^{\mathrm{pd}}(X) = P_1 X P_2$ *for some positive-definite matrices* $P_1, P_2 \in \mathbb{S}^p_{++}$.

- *The operator* M^{or} *is of one of the following two forms for orthogonal matrices* $U_1, U_2 \in \mathbb{R}^{p\times p}$*:* $\mathsf{M}^{\text{or}}(X) = U_1 X U_2$ *or* $\mathsf{M}^{\text{or}}(X) = U_1 X' U_2$.

Proof. The result follows by combining Theorem 2.2.1 with the polar decomposition. □

We refer to rank-preservers of the type M^{pd} in this corollary as *positive-definite rank-preservers* and to those of the type M^{or} as *orthogonal rank-preservers*. This corollary highlights the point that the key source of difficulty in identifying a regularizer uniquely from a factorization is due to positive-definite rank-preservers. A natural approach to address this challenge is to put a given linear map $\mathcal{L}$ into a 'canonical' form that removes the ambiguity due to positive-definite rank-preservers. In other words, we seek a distinguished subset of *normalized* linear maps with the following properties: (a) for a linear map $\mathcal{L}$, the set $\{\mathcal{L} \circ \mathsf{M}^{\text{pd}} \ : \ \mathsf{M}^{\text{pd}}$ is a positive-definite rank-preserver$\}$ intersects the collection of normalized maps at precisely one point; and (b) for any normalized linear map $\mathcal{L}$, every element of the set $\{\mathcal{L} \circ \mathsf{M}^{\text{or}} \ : \ \mathsf{M}^{\text{or}}$ is an orthogonal rank-preserver$\}$ is also normalized. The following definition possesses both of these attributes:

Definition 2.2.3 *Let* $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ *be a linear map, and let* $\mathcal{L}_i \in \mathbb{R}^{p\times p}$, $i = 1, \ldots, d$ *be the component linear functionals of* $\mathcal{L}$*. Then* $\mathcal{L}$ *is said to be* normalized *if* $\sum_{i=1}^d \mathcal{L}_i \mathcal{L}_i{}' = pI$ *and* $\sum_{i=1}^d \mathcal{L}_i{}' \mathcal{L}_i = pI$.

The utility of this definition in resolving our identifiability issue is based on a paper by Gurvits [75]. Specifically, for a generic linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$, the results in [75] imply that there exists a *unique* positive-definite rank-preserver $\mathsf{N}_{\mathcal{L}} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ so that $\mathcal{L} \circ \mathsf{N}_{\mathcal{L}}$ is normalized (see Corollary 2.2.5 in the sequel); this feature address our first requirement above. One can also check that the second requirement above is satisfied by this definition – any normalized linear map composed with any orthogonal rank-preserver is also normalized. Further, the collection of normalized maps defined above may be viewed as an affine algebraic variety specified by polynomials of degree two. One can check that any notion of normalization (specified as a real variety) that satisfies the two attributes described previously cannot be an affine space, and therefore must be specified by polynomials

of degree at least two. Consequently, our definition of normalization is in some sense also as 'simple' as possible from an algebraic perspective.[2]

In addition to satisfying these appealing properties, our notion of normalization also possesses an important computational attribute – given a (generic) linear map, a normalizing positive-definite rank-preserver for the map can be computed using the Operator Sinkhorn iterative procedure developed in [75]. Thus, the following method offers a natural approach for uniquely associating a regularizer to an equivalence class of factorizations.

Obtaining a regularizer from a linear map: Given a linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ obtained from a factorization (2.3), the unit ball of the regularizer we associate to this factorization is the image of the nuclear norm ball in $\mathbb{R}^{p\times p}$ under the linear map $\mathcal{L} \circ \mathsf{N}_{\mathcal{L}}$; here $\mathsf{N}_{\mathcal{L}}$ is the unique positive-definite rank-preserver that normalizes $\mathcal{L}$ (as discussed in the sequel in Corollary 2.2.5, such unique normalizing rank-preservers exist for generic maps $\mathcal{L}$).

The soundness of this approach follows from the fact that linear maps from equivalent factorizations produce the same regularizer. We prove a result on this point in the next section (see Proposition 2.2.6), and we also discuss algorithmic consequences of the Operator Sinkhorn scaling procedure of [75].

2.2.2 Normalizing Maps via Operator Sinkhorn Scaling

From the discussion in the preceding section, a key step in associating a unique regularizer to a collection of equivalent factorizations is to normalize a given linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$. In this section we describe how this may be accomplished by appealing to the work of Gurvits [75].

Given a linear operator $\mathsf{T} : \mathbb{S}^p \to \mathbb{S}^p$ that leaves the positive-semidefinite cone invariant, Gurvits consider the question of the existence (and computation) of positive-definite matrices $P_1, P_2 \in \mathbb{S}^p_{++}$ such that the rescaled operator $\tilde{\mathsf{T}} = (P_1 \otimes P_1) \circ \mathsf{T} \circ (P_2 \otimes P_2)$ has the property that $\tilde{\mathsf{T}}(I) = \tilde{\mathsf{T}}'(I) = I$, i.e., the identity matrix is an eigenmatrix of the rescaled operator $\tilde{\mathsf{T}}$ and its adjoint [75]. This problem is an operator analog of the classical problem of transforming entrywise square nonnegative matrices to doubly stochastic matrices by diagonal congruence scaling. This *matrix scaling* problem was originally studied by Sinkhorn [132], and he developed an iterative

[2]Note that any affine variety over the reals may be defined by polynomials of degree at most two by suitably adding extra variables; in our discussion here on normalization, we consider varieties defined without additional variables.

Algorithm 1 Normalizing a linear map via the Operator Sinkhorn iteration

Input: A linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ with component functionals $\mathcal{L}_i$, $i = 1, \ldots, d$
Require: A normalized map $\mathcal{L} \circ \mathsf{M}$ where $\mathsf{M} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ is a rank-preserver that acts via conjugation by positive-definite matrices
Algorithm: Repeat until convergence
1. $R = \sum_{i=1}^d \mathcal{L}_i \mathcal{L}_i{}'$
2. $\mathcal{L}_i \leftarrow \sqrt{p} R^{-\frac{1}{2}} \mathcal{L}_i$, $i = 1, \ldots, d$
3. $C = \sum_{i=1}^d \mathcal{L}_i{}' \mathcal{L}_i$
4. $\mathcal{L}_i \leftarrow \sqrt{p} \mathcal{L}_i C^{-\frac{1}{2}}$, $i = 1, \ldots, d$

solution technique that is known as Sinkhorn scaling. Gurvits developed an operator analog of classical Sinkhorn scaling that proceeds by alternately performing the updates $\mathsf{T} \leftarrow (\mathsf{T}(I)^{-1/2} \otimes \mathsf{T}(I)^{-1/2}) \circ \mathsf{T}$ and $\mathsf{T} \leftarrow \mathsf{T} \circ (\mathsf{T}'(I)^{-1/2} \otimes \mathsf{T}'(I)^{-1/2})$; this sequence of operations is known as the *Operator Sinkhorn iteration*. The next theorem concerning the convergence of this iterative method is proved in [75]. Following the terminology in [75], a linear operator $\mathsf{T} : \mathbb{S}^p \to \mathbb{S}^p$ is *rank-indecomposable* if it satisfies the inequality $\operatorname{rank}(\mathsf{T}(Z)) > \operatorname{rank}(Z)$ for all $Z \succeq 0$ with $1 \leq \operatorname{rank}(Z) < q$; this condition is an operator analog of a matrix being irreducible.

Theorem 2.2.4 *([75, Theorem 4.6 and 4.7]) Let $\mathsf{T} : \mathbb{S}^p \to \mathbb{S}^p$ be a rank-indecomposable linear operator. There exist* unique *positive-definite matrices $P_1, P_2 \in \mathbb{S}^p_{++}$ with $\det(P_1) = 1$ such that $\tilde{\mathsf{T}} = (P_1 \otimes P_1) \circ \mathsf{T} \circ (P_2 \otimes P_2)$ satisfies the conditions $\tilde{\mathsf{T}}(I) = \tilde{\mathsf{T}}'(I) = I$. Moreover, the Operator Sinkhorn Iteration initialized with T converges to $\tilde{\mathsf{T}}$.*

Remark. *The condition $\det(P_1) = 1$ is imposed purely to avoid the ambiguity that arises from setting $P_1 \leftarrow \alpha P_1$ and $P_2 \leftarrow \frac{1}{\alpha} P_2$ for positive scalars α. Other than this degree of freedom, there are no other positive-definite matrices that satisfy the property that the rescaled operator $\tilde{\mathsf{T}}$ in this theorem as well as its adjoint both have the identity as an eigenmatrix.*

These ideas and results are directly relevant in our context as follows. For any linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$, we may associate an operator $\mathsf{T}_{\mathcal{L}} : \mathbb{S}^p \to \mathbb{S}^p$ defined as $\mathsf{T}_{\mathcal{L}}(Z) = \frac{1}{p} \sum_{i=1}^d \mathcal{L}_i Z \mathcal{L}_i{}'$, which has the property that it leaves the positive-semidefinite cone invariant. Rescaling the operator $\mathsf{T}_{\mathcal{L}}$ via positive-definite matrices $P_1, P_2 \in \mathbb{S}^p_{++}$ to obtain $\tilde{\mathsf{T}}_{\mathcal{L}} = (P_1 \otimes P_1) \circ \mathsf{T}_{\mathcal{L}} \circ (P_2 \otimes P_2)$ corresponds to conjugating the component linear functionals $\{\mathcal{L}_i\}_{i=1}^d$ of $\mathcal{L}$ by P_1 and P_2. Consequently, rescaling

$\mathsf{T}_{\mathcal{L}}$ so that $\tilde{\mathsf{T}}_{\mathcal{L}} = (P_1 \otimes P_1) \circ \mathsf{T}_{\mathcal{L}} \circ (P_2 \otimes P_2)$ and its adjoint both have the identity as an eigenmatrix is equivalent to composing $\mathcal{L}$ by a positive-definite rank-preserver $\mathsf{N} = P_1 \otimes P_2$ so that $\mathcal{L} \circ \mathsf{N}$ is normalized. Based on this correspondence, Algorithm 1 gives a specialization of the general Operator Sinkhorn Iteration to our setting for normalizing a linear map $\mathcal{L}$.[3] We also have the following corollary to Theorem 2.2.4:

Corollary 2.2.5 *Let $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ be a linear map, and suppose* $\operatorname{rank}(\sum_{i=1}^d \mathcal{L}_i Z \mathcal{L}_i{}') > \operatorname{rank}(Z)$ *for all $Z \succeq 0$ with $1 \leq \operatorname{rank}(Z) < p$ (i.e., the operator $\mathsf{T}_{\mathcal{L}}(Z) = \frac{1}{p}\sum_{i=1}^d \mathcal{L}_i Z \mathcal{L}_i{}'$ is rank-indecomposable). There exists a* unique *positive-definite rank-preserver $\mathsf{N}_{\mathcal{L}} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ such that $\mathcal{L} \circ \mathsf{N}_{\mathcal{L}}$ is normalized. Moreover, Algorithm 1 initialized with $\mathcal{L}$ converges to $\mathcal{L} \circ \mathsf{N}_{\mathcal{L}}$.*

Proof. The existence of a positive-definite rank preserver $\mathsf{N}_{\mathcal{L}}$ as well as the convergence of Algorithm 1 follow directly from Theorem 2.2.4. We need to prove that $\mathsf{N}_{\mathcal{L}}$ is unique. Let $\tilde{\mathsf{N}}_{\mathcal{L}} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ be any positive-definite rank-preserver such that $\mathcal{L} \circ \tilde{\mathsf{N}}_{\mathcal{L}}$ is normalized. By Theorem 2.2.1, there exists positive-definite matrices $P_1, P_2, \tilde{P}_1, \tilde{P}_2$ such that $\mathsf{N}_{\mathcal{L}} = P_1 \otimes P_2$ and $\tilde{\mathsf{N}}_{\mathcal{L}} = \tilde{P}_1 \otimes \tilde{P}_2$. Without loss of generality, we may assume that $\det(P_1) = \det(\tilde{P}_1) = 1$. By Theorem 2.2.4 we have $P_1 = \tilde{P}_1$ and $P_2 = \tilde{P}_2$, and consequently that $\mathsf{N}_{\mathcal{L}} = \tilde{\mathsf{N}}_{\mathcal{L}}$. □

Generic linear maps $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ (for $d \geq 2$) satisfy the condition $\operatorname{rank}(\sum_{i=1}^d \mathcal{L}_i Z \mathcal{L}_i{}') > \operatorname{rank}(Z)$ for all $Z \succeq 0$ with $1 \leq \operatorname{rank}(Z) < p$. Therefore, this assumption in Corollary 2.2.5 is not particularly restrictive. A consequence of the uniqueness of the positive-definite rank-preserver $\mathsf{N}_{\mathcal{L}}$ in Corollary 2.2.5 is that our normalization scheme associates a unique regularizer to every collection of equivalent factorizations:

Proposition 2.2.6 *Let $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ be a linear map, and suppose* $\operatorname{rank}(\sum_{i=1}^d \mathcal{L}_i Z \mathcal{L}_i{}') > \operatorname{rank}(Z)$ *for all $Z \succeq 0$ with $1 \leq \operatorname{rank}(Z) < p$. Let $\mathsf{M} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ be any rank-preserver. Suppose $\mathsf{N}_{\mathcal{L}}$ and $\mathsf{N}_{\mathcal{L}\circ\mathsf{M}}$ are positive-definite rank-preservers such that $\mathcal{L} \circ \mathsf{N}_{\mathcal{L}}$ and $\mathcal{L} \circ \mathsf{M} \circ \mathsf{N}_{\mathcal{L}\circ\mathsf{M}}$ are normalized. Then the image of the nuclear norm ball under $\mathcal{L} \circ \mathsf{N}_{\mathcal{L}}$ is the same as it is under $\mathcal{L} \circ \mathsf{M} \circ \mathsf{N}_{\mathcal{L}\circ\mathsf{M}}$.*

[3] Algorithm 1 requires the computation of a matrix square root at every iteration. By virtue of the fact that the operator $\mathsf{T}_{\mathcal{L}}$ which we wish to rescale is completely positive, it is possible to normalize $\mathcal{L}$ using only rational matrix operations via a modified scheme known as the Rational Operator Sinkhorn iteration [75].

Remark. *Note that if the linear map $\mathcal{L}$ satisfies the property that* $\operatorname{rank}(\sum_{i=1}^{d} \mathcal{L}_i Z \mathcal{L}_i{}') > \operatorname{rank}(Z)$ *for all $Z \succeq 0$ with $1 \leq \operatorname{rank}(Z) < p$, then so does the linear map $\mathcal{L} \circ \mathsf{M}$ for any rank-preserver M.*

Proof. As $\mathsf{M}^{-1} \circ \mathsf{N}_{\mathcal{L}}$ is a rank-preserver, we can apply Corollary 2.2.2 to obtain the decomposition $\mathsf{M}^{-1} \circ \mathsf{N}_{\mathcal{L}} = \bar{\mathsf{M}}^{\mathrm{or}} \circ \bar{\mathsf{M}}^{\mathrm{pd}}$, where $\bar{\mathsf{M}}^{\mathrm{or}}$ is an orthogonal rank-preserver and $\bar{\mathsf{M}}^{\mathrm{pd}}$ is a positive-definite rank-preserver.

We claim that $\mathsf{N}_{\mathcal{L}\circ\mathsf{M}} = \mathsf{M}^{-1} \circ \mathsf{N}_{\mathcal{L}} \circ \bar{\mathsf{M}}^{\mathrm{or}'}$. First, we have $\mathsf{M}^{-1} \circ \mathsf{N}_{\mathcal{L}} \circ \bar{\mathsf{M}}^{\mathrm{or}'} = \bar{\mathsf{M}}^{\mathrm{or}} \circ \bar{\mathsf{M}}^{\mathrm{pd}} \circ \bar{\mathsf{M}}^{\mathrm{or}'}$, which implies that this operator is positive-definite. Next, we note that a linear map that is obtained by right multiplication of a normalized linear map with an orthogonal rank-preserver is also normalized, and hence the linear map $\mathcal{L} \circ \mathsf{M} \circ \mathsf{M}^{-1} \circ \mathsf{N}_{\mathcal{L}} \circ \bar{\mathsf{M}}^{\mathrm{or}'} = \mathcal{L} \circ \mathsf{N}_{\mathcal{L}} \circ \bar{\mathsf{M}}^{\mathrm{or}'}$ is normalized. By applying Corollary 2.2.5, we conclude that $\mathsf{N}_{\mathcal{L}\circ\mathsf{M}} = \mathsf{M}^{-1} \circ \mathsf{N}_{\mathcal{L}} \circ \bar{\mathsf{M}}^{\mathrm{or}'}$.

Consequently, we have $\mathcal{L} \circ \mathsf{M} \circ \mathsf{N}_{\mathcal{L}\circ\mathsf{M}} = \mathcal{L} \circ \mathsf{N}_{\mathcal{L}} \circ \bar{\mathsf{M}}^{\mathrm{or}'}$. As the nuclear norm ball is invariant under the action of the orthogonal rank-preserver $\bar{\mathsf{M}}^{\mathrm{or}'}$, it follows that the image of the nuclear norm ball under the map $\mathcal{L} \circ \mathsf{N}_{\mathcal{L}}$ is the same as it is under the map $\mathcal{L} \circ \mathsf{M} \circ \mathsf{N}_{\mathcal{L}\circ\mathsf{M}}$. □

The polynomial-time complexity of the (general) Operator Sinkhorn iterative procedure – in terms of the number of iterations required to obtain a desired accuracy to the fixed-point – has recently been established in [62]. In summary, this approach provides a computationally tractable method to normalize linear maps, and consequently to associate a unique regularizer to a collection of equivalent factorizations.

2.2.3 An Alternating Update Algorithm for Matrix Factorization

Given the resolution of the identifiability issues in the preceding two sections, we are now in a position to describe an algorithmic approach for computing a factorization (2.3) of a data matrix $Y = [\mathbf{y}^{(1)}|\cdots|\mathbf{y}^{(n)}] \in \mathbb{R}^{d\times n}$ to obtain a semidefinite regularizer that promotes the type of structure contained in Y. Specifically, given a target dimension p, our objective is to obtain a normalized linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ and a collection $\{X^{(j)}\}_{j=1}^{n}$ of low-rank matrices such that $\sum_{i=1}^{n} \|\mathbf{y}^{(j)} - \mathcal{L}(X^{(j)})\|_{\ell_2}^2$ is minimized. Our procedure is an alternating update technique that sequentially updates the low-rank $X^{(j)}$'s followed by an update of $\mathcal{L}$. We assume that our algorithm is provided with a data matrix $Y \in \mathbb{R}^{d\times n}$, a target dimension p, and an initial guess for the normalized map $\mathcal{L}$. Our method is summarized in Algorithm 3.

Algorithm 2 Obtaining a low-rank matrix near an affine space via Singular Value Projection

Input: A linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$, a point $\mathbf{y} \in \mathbb{R}^d$, a target rank r, an initial guess $X \in \mathbb{R}^{p\times p}$, and a damping parameter $\nu \in (0, 1]$
Require: A matrix $\hat{X}$ of rank at most r such that $\|\mathbf{y} - \mathcal{L}(\hat{X})\|_{\ell_2}$ is minimized, i.e., solve (2.5)
Initialization $X = 0$
Algorithm: Repeat until convergence
1. $X \leftarrow X + \nu\mathcal{L}'(\mathbf{y} - \mathcal{L}(X))$ (i.e., take a gradient step with respect to the objective of (2.5))
2. Compute top-r singular vectors and singular values of X: $U_r, V_r \in \mathbb{R}^{p\times r}$, $\Sigma_r \in \mathbb{R}^{r\times r}$
3. $X \leftarrow U_r\Sigma_r V_r'$

2.2.3.1 Updating the low-rank matrices $\{X^{(j)}\}_{j=1}^n$

In this stage a normalized linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ is fixed, and the objective is to find low-rank matrices $\{X^{(j)}\}_{j=1}^n$ such that $\mathbf{y}^{(j)} \approx \mathcal{L}(X^{(j)})$ for each $j = 1, \ldots, n$. Without the requirement that the $X^{(j)}$'s be low-rank, such linear inverse problems are ill-posed in our context as p^2 is typically taken to be larger than d. With the low-rank restriction, this problem is well-posed and it is known as the *affine rank minimization problem*. This problem is NP-hard in general [102]. However, due to its prevalence in a range of application domains [53, 116], significant efforts have been devoted towards the development of tractable heuristics that are useful in practice and that succeed on certain families of problem instances. We describe next two popular heuristics for this problem.

The first approach – originally proposed by Fazel in her thesis [53] and subsequently analyzed in [30, 116] – is based on a convex relaxation in which the rank constraint is replaced by the nuclear norm penalty, which leads to the following convex program:

$$\hat{X} = \underset{X\in\mathbb{R}^{p\times p}}{\arg\min} \quad \tfrac{1}{2}\|\mathbf{y} - \mathcal{L}(X)\|_{\ell_2}^2 + \lambda\|X\|_\star. \tag{2.4}$$

Here $\mathbf{y} \in \mathbb{R}^d$ and $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ are the problem data specifying the affine space near which we seek a low-rank solution, and the parameter $\lambda > 0$ provides a tradeoff between fidelity to the data (i.e., fit to the specified affine space) and rank of the solution $\hat{X}$. This problem is a semidefinite program and it can solved to a desired precision in polynomial-time using standard software [103, 145].

Another popular method for the affine rank minimization problem is based on directly attempting to solve the following non-convex optimization problem via alternating

projection for a specified rank $r < p$:

$$\begin{aligned} \hat{X} = \underset{X \in \mathbb{R}^{p \times p}}{\arg\min} \quad & \|\mathbf{y} - \mathcal{L}(X)\|_{\ell_2}^2 \\ \text{s.t.} \quad & \operatorname{rank}(X) \leq r. \end{aligned} \tag{2.5}$$

This problem is intractable to solve globally in general, but the heuristic described in Algorithm 2 provides an approach that provably succeeds under certain conditions [65, 82]. The utility of this method in comparison to the convex program (2.4) is that applying the procedure described in Algorithm 2 is much more tractable in large-scale settings in comparison to solving (2.4).

The analyses in [54, 65, 82, 116] rely on the map $\mathcal{L}$ satisfying the following type of restricted isometry condition introduced in [116]:

Definition 2.2.7 *Consider a linear map $\mathcal{L} : \mathbb{R}^{p \times p} \to \mathbb{R}^d$. For each $k = 1, \ldots, p$ the* restricted isometry constant of order k *is defined as the smallest $\delta_k(\mathcal{L})$ such that:*

$$1 - \delta_k(\mathcal{L}) \leq \frac{\|\mathcal{L}(X)\|_{\ell_2}^2}{\|X\|_{\ell_2}^2} \leq 1 + \delta_k(\mathcal{L})$$

for all matrices $X \in \mathbb{R}^{p \times p}$ with rank less than or equal to k.

If a linear map $\mathcal{L}$ has a small restricted isometry constant for some order k, then the affine rank minimization problem is, in some sense, well-posed when restricted to matrices of rank less than or equal to k. The results in [54, 65, 82, 116] go much further by demonstrating that if $\mathbf{y} = \mathcal{L}(X^\star) + \boldsymbol{\varepsilon}$ for $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ and with $\operatorname{rank}(X^\star) \leq r$, and if the map $\mathcal{L}$ satisfies a bound on the restricted isometry constant $\delta_{4r}(\mathcal{L})$, then both the convex program (2.4) as well as the procedure in Algorithm 2 applied to solve (2.5) provide solutions $\hat{X}$ such that $\|\hat{X} - X^\star\|_{\ell_2} \lesssim C\|\boldsymbol{\varepsilon}\|_{\ell_2}$. Due to the qualitative similarity in the performance guarantees for these approaches, either of them is appropriate as a subroutine for updating the $X^{(j)}$'s in our alternating update method for computing a factorization of a given data matrix $Y \in \mathbb{R}^{d \times n}$. Algorithm 3 is therefore stated in a general manner to retain this flexibility. In our main theoretical result in Section 2.3.3, we assume that the $X^{(j)}$'s are updated by solving (2.5) using the heuristic outlined in Algorithm 2; our analysis could equivalently be carried out by assuming that the $X^{(j)}$'s are updated by solving (2.4).

2.2.3.2 Updating the linear map $\mathcal{L}$

In this stage the low-rank matrices $\{X^{(j)}\}_{j=1}^n$ are fixed and the goal is to obtain a normalized linear map $\mathcal{L}$ such that $\sum_{i=1}^n \|\mathbf{y}^{(j)} - \mathcal{L}(X^{(j)})\|_{\ell_2}^2$ is minimized. Our

Algorithm 3 Computing a factorization via alternating updates

Input: A data matrix $Y = [\mathbf{y}^{(1)}|\cdots|\mathbf{y}^{(n)}] \in \mathbb{R}^{d\times n}$, a target dimension p, an initial guess for a normalized linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$, a target rank $r < p$

Require: A normalized linear map $\hat{\mathcal{L}} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ and a collection of matrices $\{\hat{X}^{(j)}\}_{j=1}^n$ with rank at most r such that $\sum_{i=1}^n \|\mathbf{y}^{(j)} - \hat{\mathcal{L}}(\hat{X}^{(j)})\|_{\ell_2}^2$ is minimized

Algorithm: Repeat until convergence

1.[Update $X^{(j)}$'s; $\mathcal{L}$ fixed] Obtain matrices $\{X^{(j)}\}_{j=1}^n$ of rank at most r such that $\sum_{i=1}^n \|\mathbf{y}^{(j)} - \mathcal{L}(X^{(j)})\|_{\ell_2}^2$ is minimized. This can be accomplished either via Algorithm 2 or by solving (2.4) for a suitable choice of λ.

2.[Update $\mathcal{L}$; $X^{(j)}$'s fixed] $\tilde{\mathcal{L}} \leftarrow \underset{\substack{\bar{\mathcal{L}}:\mathbb{R}^{p\times p}\to\mathbb{R}^d \\ \bar{\mathcal{L}} \text{ is a linear map}}}{\arg\min} \sum_{i=1}^n \|\mathbf{y}^{(j)} - \bar{\mathcal{L}}(X^{(j)})\|_{\ell_2}^2$

3.[Normalize $\mathcal{L}$] Normalize updated linear map from previous step using Algorithm 1.

procedure for this update consists of two steps. First we solve the following least-squares problem:

$$\tilde{\mathcal{L}} = \underset{\substack{\bar{\mathcal{L}}:\mathbb{R}^{p\times p}\to\mathbb{R}^d \\ \bar{\mathcal{L}} \text{ is a linear map}}}{\arg\min} \sum_{i=1}^n \|\mathbf{y}^{(j)} - \bar{\mathcal{L}}(X^{(j)})\|_{\ell_2}^2 \tag{2.6}$$

This problem can be solved, for example, via a pseudoinverse computation. Next, we apply the procedure described in Algorithm 1 to the updated $\tilde{\mathcal{L}}$ obtained from (2.6) in order to normalize it.

2.2.4 Comparison with Dictionary Learning

As described in Section 2.1.1, the dictionary learning literature considers the following factorization problem: given a collection of data points $\{\mathbf{y}^{(j)}\}_{j=1}^n \subset \mathbb{R}^d$ and a target dimension q, find a linear map $L : \mathbb{R}^q \to \mathbb{R}^d$ and a collection of sparse vectors $\{\mathbf{x}^{(j)}\}_{j=1}^n \subset \mathbb{R}^q$ such that $\mathbf{y}^{(j)} = L\mathbf{x}^{(j)}$ for each j. As with (2.3), the linear map L does not lead to a unique polyhedral regularizer. Specifically, for any linear sparsity-preserver $M : \mathbb{R}^q \to \mathbb{R}^q$, there is an equivalent factorization in which the linear map is LM. In parallel to Corollary 2.2.2, one can check that M is a sparsity-preserver if and only if M is a composition of a positive-definite diagonal matrix and a signed permutation matrix. Since the ℓ_1 ball is invariant under the action of a signed permutation, the main source of difficulty in obtaining a unique regularizer from a factorization is due to sparsity-preservers that are positive-definite diagonal matrices. A common convention in dictionary learning that addresses this identifiability issue is to require that each of the columns of L has unit Euclidean norm; for a generic linear map L, there is a unique positive-definite diagonal matrix D

such that LD consists of unit-norm columns. Adopting a similar reasoning as in Section 2.2.2, one can check that this normalization resolves the issue of associating a unique regularizer to an equivalence of factorizations.

The most popular approach for computing a factorization in dictionary learning is based on alternately updating the map L and the sparse vectors $\{\mathbf{x}^{(j)}\}_{j=1}^n$. For a fixed linear map L, updating the $\mathbf{x}^{(j)}$'s entails the solution of a sparse linear inverse problem for each j. That is, for each j we seek a sparse vector $\mathbf{x}^{(j)}$ in the affine space $\mathbf{y}^{(j)} = L\mathbf{x}$. Although this problem is NP-hard in general, there is significant literature on tractable heuristics that succeed under suitable conditions [31, 32, 36, 48–50]; indeed, this work predates and served as a foundation for the literature on the affine rank minimization problem. Prominent examples include the lasso [144], which is a convex relaxation approach akin to (2.4), and iterative hard thresholding [24], which is analogous to Algorithm 2. For a fixed collection $\{\mathbf{x}^{(j)}\}_{j=1}^n$, the linear map L is then updated by solving a least-squares problem followed by a rescaling of the columns so that they have unit Euclidean norm.

We note that each step in this procedure has a direct parallel to a corresponding step of Algorithm 3. In summary, our proposed approach for obtaining a semidefinite regularizer via matrix factorization is a generalization of previous methods in the dictionary learning literature for obtaining a polyhedral regularizer.

2.3 Convergence Analysis of Our Algorithm

This section describes the main theoretical result on the local convergence of our algorithm. We begin by discussing the setup and an outline of our analysis in Sections 2.3.1 and 2.3.2 respectively. The statement of our main theorem with deterministic conditions is given in Section 2.3.3, and we describe natural random ensembles that satisfy these deterministic conditions with high probability in Section 2.3.4. The proof of our theorem is discussed in Section 2.3.5.

2.3.1 Theoretical Setup

The setup underlying our main theorem is as follows. We assume that we are given a collection of data points $\{\mathbf{y}^{(j)\star}\}_{j=1}^n \subset \mathbb{R}^d$ with each $\mathbf{y}^{(j)\star} = \mathcal{L}^\star(X^{(j)\star})$, where $\mathcal{L}^\star : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ is a linear map and $\mathscr{X}^\star := \{X^{(j)\star}\}_{j=1}^n \subset \mathbb{R}^{p\times p}$ is a collection of low-rank matrices. Without loss of generality, we may take $\mathcal{L}^\star$ to be normalized and surjective. Our objective is to obtain a linear map $\hat{\mathcal{L}} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ with the property that the image of the nuclear norm ball in $\mathbb{R}^{p\times p}$ under $\mathcal{L}^\star$ is the same as it is under $\hat{\mathcal{L}}$. To this end, we seek a linear map $\hat{\mathcal{L}}$ that can be expressed as the composition of

$\mathcal{L}^\star$ with an orthogonal rank-preserver (recall that the nuclear norm ball is invariant under the action of an orthogonal rank-preserver).

As this goal is distinct from the more restrictive requirement that $\hat{\mathcal{L}}$ must equal $\mathcal{L}^\star$, we need an appropriate measure of the 'distance' of a linear map to $\mathcal{L}^\star$. A convenient approach to addressing this issue is to express a linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ in terms of $\mathcal{L}^\star$ as follows, given any linear rank-preserver $\mathsf{M} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$:

$$\mathcal{L} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}) \circ \mathsf{M}, \tag{2.7}$$

Here $\mathsf{I} \in \mathrm{End}(\mathbb{R}^{p\times p})$ is the identity map and the error term $\mathsf{E} = \mathcal{L}^{\star+} \circ (\mathcal{L} \circ \mathsf{M}^{-1} - \mathcal{L}^\star) \in \mathrm{End}(\mathbb{R}^{p\times p})$; the assumption that $\mathcal{L}^\star$ is surjective is key as $\mathcal{L}^{\star+}$ is the right-inverse of $\mathcal{L}^\star$. By varying the rank-preserver M in (2.7) the error term E changes. If there exists an *orthogonal* rank-preserver M such that the corresponding error E is small, then in some sense the image of the nuclear norm ball under $\mathcal{L}$ is close to the image under $\mathcal{L}^\star$. This observation suggests that the closeness between $\mathcal{L}$ and $\mathcal{L}^\star$ may be measured as the smallest error E that one can obtain by varying M over the set of orthogonal rank-preservers. The following result suggests that one can in fact vary M over *all* rank-preservers, provided we have the additional condition that $\mathcal{L}$ is also normalized. The additional flexibility provided by varying M over all rank-preservers is well-suited to characterizing the effects of normalization via Operator Sinkhorn scaling in our analysis, as described in the next section.

Proposition 2.3.1 *Suppose $\mathcal{L}, \mathcal{L}^\star : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ are normalized linear maps such that (i) $\mathcal{L}^\star$ satisfies the restricted isometry condition $\delta_1(\mathcal{L}^\star) \le 1/10$, and (ii) $\mathcal{L} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}) \circ \mathsf{M}$ for a linear rank-preserver M with $\|\mathsf{E}\|_{\ell_2} \le 1/(150\sqrt{p}\|\mathcal{L}^\star\|_2)$. Then there exists an orthogonal rank-preserver M^{or} such that $\|\mathsf{M}^{\mathrm{or}} - \mathsf{M}\|_2 \le 300\sqrt{p}\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2}$.*

In words, if both $\mathcal{L}$ and $\mathcal{L}^\star$ are normalized and if there exists a rank-preserver M such that $\|\mathsf{E}\|_{\ell_2}$ is small in (2.7), then M is close to an orthogonal rank-preserver[4]; in turn, this implies that the image of the nuclear norm ball under $\mathcal{L}^\star$ is close to the image of the nuclear norm ball under $\mathcal{L}$. These observations motivate the following definition as a measure of the distance between normalized linear maps $\mathcal{L}^\star, \mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ for surjective $\mathcal{L}^\star$:

$$\xi_{\mathcal{L}^\star}(\mathcal{L}) := \inf\{\|\mathsf{E}\|_{\ell_2} \mid \exists \mathsf{E} \in \mathrm{End}(\mathbb{R}^{p\times p}) \text{ and a rank-preserver } \mathsf{M} \in \mathrm{End}(\mathbb{R}^{p\times p}) \\ \text{s.t. } \mathcal{L} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}) \circ \mathsf{M}\}. \tag{2.8}$$

[4]The restricted isometry condition in Proposition 2.3.1 is a mild one; we require a stronger restricted isometry condition on $\mathcal{L}^\star$ in Theorem 2.3.5.

In Section 2.3.3, our main result gives conditions under which the sequence of normalized linear maps obtained from Algorithm 3 converges to $\mathcal{L}^\star$ in terms of the distance measure ξ.

2.3.2 An Approach for Proving a Local Convergence Result

We describe a high-level approach for proving a local convergence result, which motivates the definition of the key parameters that govern the performance of our algorithm. Our proof strategy is to demonstrate that under appropriate conditions the sequence of normalized iterates $\mathcal{L}^{(t)}$ obtained from Algorithm 3 satisfies $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)}) \leq \gamma \xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})$ for a suitable $\gamma < 1$. To bound $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)})$ with respect to $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})$, we consider each of the three steps in Algorithm 3. Fixing notation before we proceed, let $\mathcal{L}^{(t)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)}) \circ \mathsf{M}^{(t)}$ for some linear rank-preserver $\mathsf{M}^{(t)}$ and for a corresponding error term $\mathsf{E}^{(t)}$. Our objective is to show that there exists a linear rank-preserver $\mathsf{M}^{(t+1)}$ and corresponding error term $\mathsf{E}^{(t+1)}$ with $\mathcal{L}^{(t+1)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t+1)}) \circ \mathsf{M}^{(t+1)}$, so that $\|\mathsf{E}^{(t+1)}\|_{\ell_2}$ is suitably bounded above in terms of $\|\mathsf{E}^{(t)}\|_{\ell_2}$. By taking limits we obtain the desired result in terms of $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})$ and $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)})$.

The first step of Algorithm 3 involves the solution of the following optimization problem for each $j = 1, \ldots, n$:

$$\hat{X}^{(j)} = \operatorname*{arg\,min}_{X \in \mathbb{R}^{p \times p}} \left\| \mathbf{y}^{(j)\star} - \mathcal{L}^{(t)}(X) \right\|_{\ell_2}^2 \text{ s.t. } \operatorname{rank}(X) \leq r.$$

As $\mathcal{L}^{(t)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)}) \circ \mathsf{M}^{(t)}$ and as $\mathbf{y}^{(j)\star} = \mathcal{L}^\star(X^{(j)\star})$, the preceding problem can be reformulated in the following manner:

$$\mathsf{M}^{(t)}(\hat{X}^{(j)}) = \operatorname*{arg\,min}_{\tilde{X} \in \mathbb{R}^{p \times p}} \left\| \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)})(X^{(j)\star}) - \mathcal{L}^\star \circ \mathsf{E}^{(t)}(X^{(j)\star}) - \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)})(\tilde{X}) \right\|_{\ell_2}^2$$
$$\text{s.t. } \operatorname{rank}(\tilde{X}) \leq r.$$

If $\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)})$ satisfies a suitable restricted isometry condition and if $\|\mathcal{L}^\star \circ \mathsf{E}^{(t)}(X^{(j)\star})\|_{\ell_2}$ is small, then the results in [65, 82] (as described in Section 2.2.3.1) imply that $\mathsf{M}^{(t)}(\hat{X}^{(j)}) \approx X^{(j)\star}$. In other words, if $\|\mathsf{E}^{(t)}\|_{\ell_2}$ is small and if $\mathcal{L}^\star$ satisfies a restricted isometry condition, then $\mathsf{M}^{(t)}(\hat{X}^{(j)}) \approx X^{(j)\star}$; the following result states matters formally:

Proposition 2.3.2 *Let $\mathcal{L}^\star : \mathbb{R}^{p \times p} \to \mathbb{R}^d$ be a linear map such that (i) $\mathcal{L}^\star$ is normalized, and (ii) $\mathcal{L}^\star$ satisfies the restricted isometry condition $\delta_{4r}(\mathcal{L}^\star) \leq \frac{1}{20}$. Suppose $\mathcal{L} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}) \circ \mathsf{M}$ such that (i) M is a linear rank-preserver, and (ii)*

$\|\mathsf{E}\|_{\ell_2} \leq \min\{1/(50\sqrt{p}), 1/(120r^2\|\mathcal{L}^\star\|_2)\}$. *Finally, suppose* $\mathbf{y} = \mathcal{L}^\star(X^\star)$*, where* $X^\star \in \mathbb{R}^{p\times p}$ *is a rank-*r *matrix such that* $\sigma_r(X^\star) \geq \sigma_1(X^\star)/2$*, and that* $\hat{X}$ *is the optimal solution to*

$$\hat{X} = \underset{X\in\mathbb{R}^{p\times p}}{\arg\min} \ \|\mathbf{y} - \mathcal{L}(X)\|_{\ell_2}^2 \ \text{s.t. rank}(X) \leq r. \tag{2.9}$$

Then

$$\mathsf{M}(\hat{X}) = X^\star - \left[\left(\mathcal{L}^{\star\prime}_{\mathcal{T}(X^\star)}\mathcal{L}^\star_{\mathcal{T}(X^\star)}\right)^{-1}\right]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}\left(X^\star\right) + G,$$

where $\|G\|_{\ell_2} \leq 800r^{5/2}\|\mathcal{L}^\star\|_2^2\|X^\star\|_2\|\mathsf{E}\|_{\ell_2}^2$.

In this proposition, the conclusion is well-defined as the linear map $\mathcal{L}^{\star\prime}_{\mathcal{T}(X^\star)}\mathcal{L}^\star_{\mathcal{T}(X^\star)} : \mathcal{T}(X^\star) \to \mathcal{T}(X^\star)$ is invertible due to the restricted isometry condition on $\mathcal{L}^\star$ (see Lemma 2.3.9). The proof appears in Appendix A.6, and it relies primarily on the first-order optimality conditions of the problem (2.5). To ensure that the conditions required by this proposition hold, we assume in our main theorem in Section 2.3.3 that $\mathcal{L}^\star$ satisfies the restricted isometry property for rank-r matrices and that the initial guess $\mathcal{L}^{(0)}$ that is supplied to Algorithm 3 is such that $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)})$ is small (with a sufficiently good initial guess and by an inductive hypothesis, we have that there exists an error term $\mathsf{E}^{(t)}$ at iteration t such that $\|\mathsf{E}^{(t)}\|_{\ell_2}$ is small).

The second step of Algorithm 3 entails the solution of a least-squares problem. To describe the implications of this step in detail, we consider the linear maps $\mathtt{X}^\star : \mathbf{z} \mapsto \sum_{j=1}^n X^{(j)^\star}\mathbf{z}_j$ and $\hat{\mathtt{X}} : \mathbf{z} \mapsto \sum_{j=1}^n \hat{X}^{(j)}\mathbf{z}_j$ from $\mathbb{R}^n$ to $\mathbb{R}^{p\times p}$. With this notation, the second step of Algorithm 3 results in the linear map $\mathcal{L}^{(t)}$ being updated as follows:

$$\tilde{\mathcal{L}}^{(t+1)} = \mathcal{L}^\star \circ \mathtt{X}^\star \circ \hat{\mathtt{X}}^+. \tag{2.10}$$

In order for the normalized version of $\tilde{\mathcal{L}}^{(t+1)}$ to be close to $\mathcal{L}^\star$ (in terms of the distance measure ξ), we require a deeper understanding of the structure of $\mathtt{X}^\star \circ \hat{\mathtt{X}}^+$, which is the focus of the next proposition. This result relies on the set $\mathscr{X}^\star$ being suitably isotropic, as characterized by the quantities $\Delta(\mathscr{X}^\star)$ and $\Lambda(\mathscr{X}^\star)$.

Proposition 2.3.3 *Let* $\{A^{(j)}\}_{j=1}^n \subset \mathbb{R}^{p\times p}$ *and* $\{B^{(j)}\}_{j=1}^n \subset \mathbb{R}^{p\times p}$ *be two collections of matrices, and let* $\mathtt{A} : \mathbf{z} \mapsto \sum_{j=1}^n A^{(j)}\mathbf{z}_j$ *and* $\mathtt{B} : \mathbf{z} \mapsto \sum_{j=1}^n B^{(j)}\mathbf{z}_j$ *be linear maps from* $\mathbb{R}^n$ *to* $\mathbb{R}^{p\times p}$ *associated to these ensembles. Let* $\mathsf{Q} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ *be any invertible linear operator and denote* $\omega = \max_j \left\|\mathsf{Q}(B^{(j)}) - A^{(j)}\right\|_{\ell_2}$*. If* $\omega \leq \dfrac{\sqrt{\Lambda\left(\{A^{(j)}\}_{j=1}^n\right)}}{20}$ *and if*

$\frac{\Delta\left(\{A^{(j)}\}_{j=1}^n\right)}{\Lambda\left(\{A^{(j)}\}_{j=1}^n\right)} \leq \frac{1}{6}$, *then*

$$\mathtt{A} \circ \mathtt{B}^+ = \left(\mathsf{I} - \frac{1}{n\Lambda\left(\{A^{(j)}\}_{j=1}^n\right)} \sum_{j=1}^n \left(\mathsf{Q}(B^{(j)}) - A^{(j)} \right) \boxtimes A^{(j)} + \mathsf{F} \right) \circ \mathsf{Q}, \tag{2.11}$$

where $\|\mathsf{F}\|_{\ell_2} \leq 20p\frac{\omega^2}{\Lambda\left(\{A^{(j)}\}_{j=1}^n\right)} + 2p\frac{\omega\Delta\left(\{A^{(j)}\}_{j=1}^n\right)}{\Lambda\left(\{A^{(j)}\}_{j=1}^n\right)^{3/2}}$.

The proof of this proposition appears in Appendix A.7, and it consists of two key elements. First, as ω is bounded, the operator $\mathtt{A} \circ \mathtt{B}^+$ may be approximated as $\mathtt{A} \circ \mathtt{A}^+ \circ \mathsf{Q}$. Second, as the set $\{A^{(j)}\}_{j=1}^n$ is near-isotropic based on the assumptions involving $\Delta(\{A^{(j)}\}_{j=1}^n)$ and $\Lambda(\{A^{(j)}\}_{j=1}^n)$, one can show that $\mathtt{A} \circ \mathtt{A}^+$ can be expanded suitably around the identity map I. In the context of our analysis, we apply the conclusions of Proposition 2.3.3 with the choice of $A^{(j)} = X^{(j)\star}$, $B^{(j)} = \hat{X}^{(j)}$, and $\mathsf{Q} = \mathsf{M}^{(t)}$.

The final step of our analysis is to consider the effect of normalization on the map $\tilde{\mathcal{L}}^{(t)}$ in (2.10). Denoting the positive-definite rank-preserver that normalizes $\tilde{\mathcal{L}}^{(t+1)}$ by $\mathsf{N}_{\tilde{\mathcal{L}}^{(t+1)}}$, we have from Propositions 2.3.2 and 2.3.3 that the normalized map $\mathcal{L}^{(t+1)}$ obtained after the application of the Operator Sinkhorn iterative procedure to $\tilde{\mathcal{L}}^{(t+1)}$ can be expressed as:

$$\mathcal{L}^{(t+1)} = \mathcal{L}^\star \circ \left(\mathsf{I} - \frac{1}{n\Lambda(\mathscr{X}^\star)} \sum_{j=1}^n \left(\mathsf{M}^{(t)}(\hat{X}^{(j)}) - X^{(j)\star} \right) \boxtimes X^{(j)\star} + \mathsf{F} \right) \circ \mathsf{M}^{(t)} \circ \mathsf{N}_{\tilde{\mathcal{L}}^{(t+1)}},$$

where $\mathsf{F} \in \mathrm{End}(\mathbb{R}^{p\times p})$ is suitably bounded. As $\mathsf{M}^{(t)}$ and $\mathsf{N}_{\tilde{L}^{(t+1)}}$ are both rank-preservers, we need to prove that the expression within parentheses $\mathsf{I} - \frac{1}{n\Lambda(\mathscr{X}^\star)} \sum_{j=1}^n (\mathsf{M}^{(t)}(\hat{X}^{(j)}) - X^{(j)\star}) \boxtimes X^{(j)\star} + \mathsf{F}$ is well-approximated as a rank-preserver so that $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)})$ is suitably controlled. To make progress on this front, we note that $\mathsf{I} = I \otimes I$ is a rank-preserver. Therefore, if $-\frac{1}{n\Lambda(\mathscr{X}^\star)} \sum_{j=1}^n (\mathsf{M}^{(t)}(\hat{X}^{(j)}) - X^{(j)\star}) \boxtimes X^{(j)\star} + \mathsf{F}$ is small, a natural approach to characterizing how close $\mathsf{I} - \frac{1}{n\Lambda(\mathscr{X}^\star)} \sum_{j=1}^n (\mathsf{M}^{(t)}(\hat{X}^{(j)}) - X^{(j)\star}) \boxtimes X^{(j)\star} + \mathsf{F}$ is to a rank-preserver is to express this quantity in terms of the following *tangent space* at I with respect to the set of rank-preservers acting on the space of $p \times p$ matrices:

$$\mathcal{W} = \mathrm{span}\{I \otimes W_1 + W_2 \otimes I \mid W_1, W_2 \in \mathbb{R}^{p\times p}\} \tag{2.12}$$

The next result gives such an expression.

Proposition 2.3.4 *Suppose* $\mathsf{D} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ *is a linear operator such that* $\|\mathsf{D}\|_{\ell_2} \leq 1/10$ *and* $\mathsf{I} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ *is the identity operator. Then we have that*

$$\mathsf{I} + \mathsf{D} = (\mathsf{I} + \mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D}) + \mathsf{H}) \circ \mathsf{W}$$

where $\mathsf{H} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ *is a linear operator such that* $\|\mathsf{H}\|_{\ell_2} \leq 5\|\mathsf{D}\|_{\ell_2}^2/\sqrt{p}$ *and* $\mathsf{W} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ *is a linear rank-preserver such that* $\|\mathsf{W} - \mathsf{I}\|_2 \leq 3\|\mathsf{D}\|_{\ell_2}/\sqrt{p}$. *Here, the space* $\mathcal{W}$ *is as defined in* (2.12).

The proof of this proposition appears in Appendix A.8. As detailed in the proof of Theorem 2.3.5 in Section 2.3.5, one can combine the preceding three results along with the observation that $c\ \mathcal{P}_{\mathcal{T}(X^\star)} \preceq [(\mathcal{L}^\star_{\mathcal{T}(X^\star)}{}'\mathcal{L}^\star_{\mathcal{T}(X^\star)})^{-1}]_{\mathbb{R}^{p\times p}} \preceq \tilde{c}\ \mathcal{P}_{\mathcal{T}(X^\star)}$ for suitable constants $c, \tilde{c} > 0$ (from Lemma 2.3.9 in Section 2.3.5 based on $\mathcal{L}^\star$ satisfying a suitable restricted isometry condition) to conclude that there exists an error term $\mathsf{E}^{(t+1)}$ at iteration $t+1$ (corresponding to the error term $\mathsf{E}^{(t)}$ at iteration t that we fixed at the beginning of this argument) such that

$$\begin{aligned}\mathsf{E}^{(t+1)} =& \mathcal{P}_{\mathcal{W}^\perp} \circ \left[\frac{1}{n\Lambda(\mathscr{X}^\star)} \sum_{j=1}^{n} \left(X^{(j)\star} \boxtimes X^{(j)\star}\right) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)\star})}\right] (\mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}^{(t)}) \\ &+ \mathcal{P}_{\mathcal{W}^\perp}(\mathsf{F}) + O(\|\mathsf{E}^{(t)}\|_{\ell_2}^2).\end{aligned} \tag{2.13}$$

Thus, there are two 'significant' terms in this expression that govern the size of $\|\mathsf{E}^{(t+1)}\|_{\ell_2}$. To control the first term, we require a bound on the following operator norm:

$$\Omega(\mathscr{X}^\star) := \left\| \mathcal{P}_{\mathcal{W}^\perp} \circ \left[\frac{1}{n} \sum_{j=1}^{n} \left(X^{(j)\star} \boxtimes X^{(j)\star}\right) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)\star})}\right] \right\|_2 . \tag{2.14}$$

Note that this operator belongs to $\mathrm{End}(\mathrm{End}(\mathbb{R}^{p\times p}))$. In Section 2.3.5 we show that the first significant term in (2.13) is bounded as $\frac{2\|\mathcal{L}^\star\|_2^2\Omega(\mathscr{X}^\star)}{\Lambda(\mathscr{X}^\star)}\|\mathsf{E}^{(t)}\|_{\ell_2}$. For the second term in (2.13), we show in Section 2.3.5 that $\|\mathsf{F}\|_{\ell_2} \lesssim \frac{p^2\|\mathcal{L}^\star\|_2\Delta(\mathscr{X}^\star)}{\Lambda(\mathscr{X}^\star)}\|\mathsf{E}^{(t)}\|_{\ell_2}$ based on a bound on $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)})$ on the initial guess. Consequently, two of the key assumptions in Theorem 2.3.5 concern bounds on the quantities $\frac{\Omega(\mathscr{X}^\star)}{\Lambda(\mathscr{X}^\star)}$ and $\frac{\Delta(\mathscr{X}^\star)}{\Lambda(\mathscr{X}^\star)}$.

We note that the Operator Sinkhorn scaling procedure for normalization is crucial in our algorithm. Aside from addressing the identifiability issues as discussed in Section 2.2.1, the incorporation of this method also plays an important role in the convergence of Algorithm 3. Specifically, if we do not apply this procedure in each iteration of Algorithm 3, then the estimate of $\mathcal{L}^\star$ at the end of iteration $t+1$ would be

$\tilde{\mathcal{L}}^{(t+1)}$ from (2.10). In analyzing how close the image of the nuclear norm ball under $\tilde{\mathcal{L}}^{(t+1)}$ is to the image of the nuclear norm ball under $\mathcal{L}^\star$, we would need to consider how close $\mathtt{X}^\star \circ \hat{\mathtt{X}}^+$ is to an *orthogonal* rank-preserver as opposed to an arbitrary rank preserver; in particular, we cannot apply Proposition 2.3.1 as $\tilde{\mathcal{L}}^{(t+1)}$ is not normalized. In analogy to the discussion preceding Proposition 2.3.4 and by noting that $\mathsf{I} = I \otimes I$ is an orthogonal rank-preserver, we could attempt to express $\mathtt{X}^\star \circ \hat{\mathtt{X}}^+$ in terms of the following tangent space at I with respect to the set of orthogonal rank-preservers:

$$\mathcal{S} = \text{span}\{I \otimes S_1 + S_2 \otimes I \ : \ S_1, S_2 \in \mathbb{R}^{p\times p} \text{ and skew-symmetric}\}. \tag{2.15}$$

Following similar reasoning as in the preceding paragraph, the convergence of our algorithm without normalization would be governed by $\|\mathcal{P}_{\mathcal{S}^\perp} \circ [\frac{1}{n}\sum_{j=1}^n (X^{(j)\star} \boxtimes X^{(j)\star}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)\star})}]\|_2$. This operator norm is, in general, much larger than the quantity $\Omega(\mathscr{X}^\star)$ defined in (2.14) as $\mathcal{S} \subset \mathcal{W}$, which can in turn affect the convergence of our algorithm. In particular, for a natural random ensemble $\mathscr{X}^\star$ of low-rank matrices described in Proposition 2.3.8 in Section 2.3.4, the condition on $\Omega(\mathscr{X}^\star)$ in Theorem 2.3.5 is satisfied while the analogous condition on $\|\mathcal{P}_{\mathcal{S}^\perp} \circ [\frac{1}{n}\sum_{j=1}^n (X^{(j)\star} \boxtimes X^{(j)\star}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)\star})}]\|_2$ is violated (both of these conclusions hold with high probability), thus highlighting the importance of the inclusion of the normalization step for the convergence of our method; see the remarks following Proposition 2.3.8 for details.

2.3.3 Main Result

The following theorem gives the main result concerning the local convergence of our algorithm:

Theorem 2.3.5 *Let* $\mathbf{y}^{(j)} = \mathcal{L}^\star(X^{(j)\star})$, $j = 1, \ldots, n$, *where* $\mathcal{L}^\star : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ *is a linear map and* $\mathscr{X}^\star := \{X^{(j)\star}\}_{j=1}^n \subset \mathbb{R}^{p\times p}$. *Suppose the collection* $\mathscr{X}^\star$ *satisfies the following conditions:*

1. *There exists* $r < p$ *and* $s > 0$ *such that* $\text{rank}(X^{(j)\star}) = r$ *and* $s \geq \sigma_1(X^{(j)\star}) \geq \sigma_r(X^{(j)\star}) \geq s/2$ *for each* $j = 1, \ldots, n$;
2. $\frac{\Omega(\mathscr{X}^\star)}{\Lambda(\mathscr{X}^\star)} \leq \frac{d}{40p^2}$; *and*
3. $\frac{\Delta(\mathscr{X}^\star)}{\Lambda(\mathscr{X}^\star)} \leq \frac{\sqrt{d}}{100p^3}$.

Suppose the linear map $\mathcal{L}^\star : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ *satisfies the following conditions:*

1. *$\mathcal{L}^\star$ satisfies the restricted isometry condition $\delta_{4r}(\mathcal{L}^\star) \leq \frac{1}{20}$, where r is the rank of each $X^{(j)\star}$;*

2. *$\mathcal{L}^\star$ is normalized and surjective; and*

3. *$\|\mathcal{L}^\star\|_2^2 \leq \frac{5p^2}{d}$.*

If we supply Algorithm 3 with a normalized initial guess $\mathcal{L}^{(0)} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ with $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)}) < \frac{1}{20000p^{7/2}r^2\|\mathcal{L}^\star\|_2^2}$, then the sequence $\{\mathcal{L}^{(t)}\}$ produced by the algorithm satisfies $\limsup_{t\to\infty} \frac{\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)})}{\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})} \leq 2\|\mathcal{L}^\star\|_2^2 \frac{\Omega(\mathscr{X}^\star)}{\Lambda(\mathscr{X}^\star)} + 10p^2\|\mathcal{L}^\star\|_2 \frac{\Delta(\mathscr{X}^\star)}{\Lambda(\mathscr{X}^\star)} < 1$. In other words, $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)}) \to 0$ with the rate of convergence bounded above by $2\|\mathcal{L}^\star\|_2^2 \frac{\Omega(\mathscr{X}^\star)}{\Lambda(\mathscr{X}^\star)} + 10p^2\|\mathcal{L}^\star\|_2 \frac{\Delta(\mathscr{X}^\star)}{\Lambda(\mathscr{X}^\star)}$. We assume here that Step 1 of Algorithm 3 is computed via Algorithm 2.

Remark. (*i*) *In this result the assumption that Step* 1 *of Algorithm 3 is computed via Algorithm 2 is made for the sake of concreteness. A similar result and proof are possible if Step* 1 *of Algorithm 3 is instead computed by solving* (2.4) *for a suitable choice of the regularization parameter.* (*ii*) *In conjunction with Proposition 2.3.1, this result implies that we obtain a linear map $\hat{\mathcal{L}}$ upon convergence of our algorithm such that the image of the nuclear norm ball in $\mathbb{R}^{p\times p}$ under $\hat{\mathcal{L}}$ is the same as it is under $\mathcal{L}^\star$.*

The proof of this theorem is given in Section 2.3.5. In words, our result states that under a restricted isometry condition on the linear map $\mathcal{L}^\star$ and an isotropy condition on the low-rank matrices $\{X^{(j)\star}\}_{j=1}^n$, Algorithm 3 is locally linearly convergent to the appropriate semidefinite-representable regularizer that promotes the type of structure contained in the data $\{\mathcal{L}^\star(X^{(j)\star})\}_{j=1}^n$. The restricted isometry condition on $\mathcal{L}^\star$ ensures that the geometry of the set of points $\{X^{(j)\star}\}_{j=1}^n$ in $\mathbb{R}^{p\times p}$ is (approximately) preserved in the lower-dimensional space $\mathbb{R}^d$. The isotropy condition on the collection $\{X^{(j)\star}\}_{j=1}^n$ ensures that we have observations that lie on most of the low-dimensional faces of the regularizer, which gives us sufficient information to reconstruct the regularizer.

Results of this type have previously been obtained in the classical dictionary learning literature [2, 9], although our analysis is more challenging in comparison to this prior work for two reasons. First, two nearby sparse vectors with the same number of nonzero entries have the same support, while two nearby low-rank matrices with the same rank have different row/column spaces; geometrically, this translates to

the point that two nearby sparse vectors have the same tangent space with respect to a suitably defined variety of sparse vectors, while two nearby low-rank matrices generically have different tangent spaces with respect to an appropriate variety of low-rank matrices. Second (and more significant), the normalization step in classical dictionary learning is simple – corresponding to scaling the columns of a matrix to have unit Euclidean norm, as discussed in Section 2.2.4 – while the normalization step in our setting based on Operator Sinkhorn scaling is substantially more complicated. Indeed, one of the key aspects of our analysis is the relation between the stability properties of Operator Sinkhorn scaling and the tangent spaces to varieties of low-rank matrices, as is evident from the appearance of the parameter $\Omega(\{X^{(j)\star}\}_{j=1}^{n})$ in Theorem 2.3.5.

The distance measure $\xi_{\mathcal{L}^\star}$ that appears in Theorem 2.3.5 is defined up to an equivalence relation, and with respect to the linear map $\mathcal{L}^\star$ to which we do not have access. In practice, it is useful to have a stopping criterion that only depends on the sequence of iterates. To this end, the next result states that under the same conditions as in Theorem 2.3.5, the sequence of iterates $\{\mathcal{L}^{(t)}\}$ obtained from our algorithm also converges (the limit point is generically different from $\mathcal{L}^\star$, although they specify the same regularizer):

Proposition 2.3.6 *Under the same setup and assumptions as in Theorem 2.3.5, the sequence of iterates $\{\mathcal{L}^{(t)}\}$ obtained from our algorithm is a Cauchy sequence.*

This result is proved in Appendix A.9.

Extension to the noisy case. In practice the data points $\mathbf{y}^{(j)}$ may be corrupted by noise, and it is of interest to investigate if our algorithm is robust to noise. One can extend our analysis to demonstrate the robustness of our algorithm in a stylized setting in which the data points $\mathbf{y}^{(j)}$ in Theorem 2.3.5 are corrupted by additive noise. Briefly, such an extension comprises two key steps. First, one can show that there exists a normalized linear map $\check{\mathcal{L}}$ that is close to $\mathcal{L}^\star$ (up to composition by an orthogonal rank-preserver), and which is a fixed-point of our algorithm. The key ingredient in demonstrating this is to prove that each iteration of our algorithm is *contractive* in a neighborhood of $\mathcal{L}^\star$ and to appeal to a suitable fixed-point theorem. The proximity of the regularizer defined by $\check{\mathcal{L}}$ to the regularizer defined by $\mathcal{L}^\star$ is determined by the radius of contraction, which depends linearly (under the conditions of Theorem 2.3.5) on the size of the noise corrupting the measurements

and inverse-polynomially on the size of the data set. Second, one can show that our algorithm is locally linearly convergent to $\check{\mathcal{L}}$ (up to composition by an orthogonal rank-preserver). This step essentially follows the same sequence of arguments as in the proof of Theorem 2.3.5, and it relies on the radius of contraction from the first step being smaller than the basin of attraction defined in Theorem 2.3.5; this is true as long as the noise corruption is suitably small and the number of data points is sufficiently large.

2.3.4 Ensembles Satisfying the Conditions of Theorem 2.3.5

Theorem 2.3.5 gives deterministic conditions on the underlying data under which our algorithm recovers the correct regularizer. In this section we demonstrate that these conditions are in fact satisfied with high probability by certain natural random ensembles. Our first result states that random Gaussian linear maps upon normalization satisfy the requirements on the linear map in Theorem 2.3.5:

Proposition 2.3.7 *Let $\tilde{\mathcal{L}} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ be a linear map in which each of the d component linear functionals are specified by matrices $\tilde{\mathcal{L}}_i \in \mathbb{R}^{p\times p}$ with i.i.d random Gaussian entries with mean zero and variance $1/d$. Let $\mathcal{L}$ represent a normalized map obtained by composing $\tilde{\mathcal{L}}$ with a positive-definite rank-preserver. Fix any $\delta < 1$. Then there exist positive constants c_1, c_2, c_3 depending only on δ such that if $d \geq c_1 rp$, then (i) $\delta_{4r}(\mathcal{L}) \leq \delta$ and (ii) $\|\mathcal{L}\|_2 \leq \sqrt{\frac{5p^2}{d}}$ with probability greater than $1 - c_2 \exp(-c_3 d)$.*

The proof of this result is given in Appendix A.4. As shown in [28] random Gaussian linear maps from $\mathbb{R}^{p\times p}$ to $\mathbb{R}^d$ satisfy the restricted isometry property for rank-$4r$ matrices if $d \gtrsim rp$ (and this bound is tight). Our result shows that under the same scaling assumption on d, 'most' linear maps satisfy the more restrictive requirements of Theorem 2.3.5. Next we consider families of random low-rank matrices:

Proposition 2.3.8 *Let $\mathscr{X} := \{X^{(j)}\}_{j=1}^n$ be an ensemble of matrices generated as $X^{(j)} = \sum_{i=1}^r s_i^{(j)} \mathbf{u}_i^{(j)} \mathbf{v}_i^{(j)\prime}$ with each $U^{(j)} = [\mathbf{u}_1^{(j)} | \ldots | \mathbf{u}_r^{(j)}], V^{(j)} = [\mathbf{v}_1^{(j)} | \ldots | \mathbf{v}_r^{(j)}] \in \mathbb{R}^{p\times r}$ being drawn independently from the Haar measure on $p \times r$ matrices with orthonormal columns, and each $s_i^{(j)}$ being drawn independently from $\mathcal{D}$, where $\mathcal{D}$ is any distribution supported on $[s/2, s]$ for some $s > 0$. Then for any $0 < t_1 \leq 1/4$ and $0 < t_2$, the conditions (i) $\frac{\Delta(\mathscr{X})}{\Lambda(\mathscr{X})} \leq t_1$ and (ii) $\frac{\Omega(\mathscr{X})}{\Lambda(\mathscr{X})} \leq 80\frac{r}{p} + t_2$, are satisfied*

with probability greater than $1 - 2p\exp(-\frac{nt_1^2}{200p^4}) - p\exp(-\frac{nt_2^2}{200p^4})$. *In particular, the requirements in Theorem 2.3.5 for* $d \gtrsim rp$ *are satisfied with high probability by the ensemble* $\mathscr{X}$ *provided* $n \gtrsim \frac{p^{10}}{d}$.

Considering the requirements of Theorem 2.3.5 in the regime $d \gtrsim rp$ is not restrictive as this condition is necessary for the restricted isometry assumptions of Theorem 2.3.5 on $\mathcal{L}^\star$ to hold. The proof of this result is given in Appendix A.2. Thus, in some sense, 'most' (sufficiently large) sets of low-rank matrices satisfy the requirements of Theorem 2.3.5. We also note that for a collection of low-rank matrices $\mathscr{X}$ generated according to the ensemble in this proposition, the ratio $\frac{\Delta(\mathscr{X})}{\Lambda(\mathscr{X})} \to 0$ as $n \to \infty$, while one can show that the ratio $\frac{\Omega(\mathscr{X})}{\Lambda(\mathscr{X})} \asymp \frac{r}{p}$ as $n \to \infty$. Based on Theorem 2.3.5, this observation implies that for data generated according to the ensemble in Proposition 2.3.8, the rate of convergence of Algorithm 3 improves with an increase in the amount of data, but only up to a certain point beyond which the convergence rate plateaus. We illustrate this property with a numerical experiment in Section 2.4.1.

Remark. *It is critical in the preceding result that we project onto the orthogonal complement of the subspace* $\mathcal{W}$ *from* (2.14) *in the definition of* $\Omega(\mathscr{X})$. *For a set of low-rank matrices* $\mathscr{X}$ *drawn from the same ensemble as in Proposition 2.3.8, one can show that* $\|\mathcal{P}_{\mathcal{S}^\perp} \circ \frac{1}{n}\sum_{j=1}^n (X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})}\|_2 > c\Lambda(\mathscr{X})$ *for a constant* $c > 0$ *with high probability, where the subspace* $\mathcal{S}$ *is defined in* (2.15). *In the context of the discussion at the end of the preceding section, we have that the conditions of Theorem 2.3.5 are violated if we do not incorporate the normalization step via Operator Sinkhorn scaling, which in turn impacts the convergence of our algorithm.*

2.3.5 Proof of Theorem 2.3.5

Before giving a proof of Theorem 2.3.5, we state two relevant lemmas that are proved in Appendix A.1.

Lemma 2.3.9 *Suppose a linear map* $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ *satisfies the restricted isometry condition* $\delta_{2r}(\mathcal{L}) < 1$. *For any* $\mathcal{T} := \mathcal{T}(X)$ *with* $X \in \mathbb{R}^{p\times p}$ *and* $\operatorname{rank}(X) \le r$, *we have that* (i) $1 - \delta_{2r} \le \lambda_{\min}(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}}) \le \lambda_{\max}(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}}) \le 1 + \delta_{2r}$, (ii) $\|(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}})^{-1}\|_2 = \|[(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}}\|_2 \le \frac{1}{1-\delta_{2r}}$, (iii) $\|\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}\|_2 \le \sqrt{1+\delta_{2r}}\|\mathcal{L}\|_2$, *and* (iv) $\|[(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}'\mathcal{L}\|_2 \le \frac{\sqrt{1+\delta_{2r}}}{1-\delta_{2r}}\|\mathcal{L}\|_2$. *Here* $\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}} : \mathcal{T} \to \mathcal{T}$ *is a self-adjoint linear map.*

Lemma 2.3.10 *Let $\mathscr{X} := \{X^{(j)}\}_{j=1}^n \subset \mathbb{R}^{p\times p}$ be a collection of matrices, and let $s_{\min} := \min_j \|X^{(j)}\|_{\ell_2}^2$ and $s_{\max} := \max_j \|X^{(j)}\|_{\ell_2}^2$. Then $s_{\min}/p^2 - \Delta(\mathscr{X}) \leq \Lambda(\mathscr{X}) \leq s_{\max}/p^2 + \Delta(\mathscr{X})$.*

Proof of Theorem 2.3.5. To simplify the presentation of our proof we define the following quantities $\alpha_0 := 20000p^{7/2}r^2\|\mathcal{L}^\star\|_2^2$, $\alpha_1 := 800r^{5/2}\|\mathcal{L}^\star\|_2^2$, $\alpha_2 := 2\sqrt{r}\|\mathcal{L}^\star\|_2$, $\alpha_3 := 10p^2\|\mathcal{L}^\star\|_2$, $\alpha_4 := 5(p^2/\sqrt{r})\alpha_1$, $\alpha_5 := 100p^3\alpha_2^2$, $\alpha_6 := 5(p^2/\sqrt{r})\alpha_2$, and $\alpha_7 := \alpha_3 + \alpha_6/6 + 1/4$. The specific interpretation of these quantities is not essential to the proof – the pertinent detail is that they only depend on $p, r, \|\mathcal{L}^\star\|_2$.

To simplify notation in the proof we denote $\Delta := \Delta(\mathscr{X})$, $\Lambda := \Lambda(\mathscr{X})$, and $\Omega := \Omega(\mathscr{X})$. In addition we also denote $\mathcal{T}^{(j)} := \mathcal{T}(X^{(j)\star})$. Our proof proceeds by establishing the following assertion. Suppose that the t-th iterate $\mathcal{L}^{(t)}$ is such that $\mathcal{L}^{(t)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)}) \circ \mathsf{M}^{(t)}$, where $\mathsf{M}^{(t)}$ is a rank-preserver, and $\mathsf{E}^{(t)}$ is a linear operator that satisfies $\|\mathsf{E}^{(t)}\|_{\ell_2} < 1/\alpha_0$. Then the $t+1$-th iterate is of the form $\mathcal{L}^{(t+1)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t+1)}) \circ \mathsf{M}^{(t+1)}$ for some rank-preserver $\mathsf{M}^{(t+1)}$, and some linear operator $\mathsf{E}^{(t+1)}$ that satisfies

$$\|\mathsf{E}^{(t+1)}\|_{\ell_2} \leq \gamma_0\|\mathsf{E}^{(t)}\|_{\ell_2} + \gamma_1\|\mathsf{E}^{(t)}\|_{\ell_2}^2, \tag{2.16}$$

where $\gamma_0 = 2\|\mathcal{L}^\star\|_2^2(\Omega/\Lambda) + \alpha_6(\Delta/\Lambda)$, and $\gamma_1 = \alpha_4 + \alpha_5 + 5\alpha_7^2/\sqrt{p}$.

Before we prove this assertion, we note how it allows us to conclude the result. By taking the infimum over $\mathsf{E}^{(t)}$ on the right hand side of (2.16) and by noting that $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)}) \leq \|\mathsf{E}^{(t+1)}\|_{\ell_2}$, we have

$$\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)}) \leq \gamma_0\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)}) + \gamma_1\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})^2. \tag{2.17}$$

One can check based on the initial assumption on $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)})$ that $\gamma := \gamma_0 + \gamma_1\xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)}) < 1$. By employing an inductive argument one can establish that $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)}) \leq \gamma\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})$. Thus $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)}) \leq \gamma^t\xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)}) \to 0$ as $t \to \infty$. By dividing the expression in (2.17) throughout by $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})$, and subsequently taking the limit $t \to \infty$, we obtain the asymptotic rate of convergence

$$\limsup_{t\to\infty} \frac{\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t+1)})}{\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})} \leq \limsup_{t\to\infty}\big(\gamma_0 + \gamma_1\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)})\big) = \gamma_0.$$

We proceed to prove the assertion.

[Applying Proposition 2.3.2]: Since $\|\mathsf{E}^{(t)}\|_{\ell_2} \leq \min\{1/(50\sqrt{p}), 1/(120r^2\|\mathcal{L}^\star\|_2)\}$, by applying Proposition 2.3.2 with the choice of $X^\star = X^{(j)\star}$, $\mathsf{E} = \mathsf{E}^{(t)}$, $\mathsf{M} = \mathsf{M}^{(t)}$,

and $\mathcal{L}^\star$, we have for each $j = 1, \ldots, n$ that

$$\mathsf{M}^{(t)}(\hat{X}^{(j)}) - X^{(j)\star} = -\left[[(\mathcal{L}^{\star\prime}_{\mathcal{T}^{(j)}}\mathcal{L}^{\star}_{\mathcal{T}^{(j)}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^{\star} \circ \mathsf{E}^{(t)}\right]\left(X^{(j)\star}\right) + G^{(j)}, \tag{2.18}$$

where $G^{(j)}$ is a matrix that satisfies $\|G^{(j)}\|_{\ell_2} \leq \alpha_1 \|X^{(j)\star}\|_2 \|\mathsf{E}^{(t)}\|^2_{\ell_2}$.

[Applying Proposition 2.3.3]: The next step is to apply Proposition 2.3.3 to the collections of matrices $\{X^{(j)\star}\}_{j=1}^n$ and $\{\hat{X}^{(j)}\}_{j=1}^n$. Let $\mathsf{X}^\star, \hat{\mathsf{X}}$ denote the linear maps $\mathsf{X}^\star : \mathbf{z} \mapsto \sum_{j=1}^n X^{(j)\star}\mathbf{z}_j$, $\hat{\mathsf{X}} : \mathbf{z} \mapsto \sum_{j=1}^n \hat{X}^{(j)}\mathbf{z}_j$. First note that $\alpha_1\|\mathsf{E}^{(t)}\|_{\ell_2} \leq \alpha_1/\alpha_0 \leq \sqrt{r}\|\mathcal{L}^\star\|_2$. Second from the assumptions we have $\Delta/\Lambda \leq 1/21$. Hence by Lemma 2.3.10 we have $\Lambda \leq s^2 r/p^2 + \Delta \leq s^2 r/p^2 + \Lambda/21$. It follows that $\Delta \leq s^2 r/(20p^2)$, and thus by Lemma 2.3.10 we have $\Lambda \geq s^2 r/(5p^2)$. Third by applying these inequalities and Lemma 2.3.9 to (2.18) we have $\|\mathsf{M}^{(t)}(\hat{X}^{(j)}) - X^{(j)\star}\|_{\ell_2} \leq ((\sqrt{1+\delta_{4r}})/(1-\delta_{4r}))\|\mathcal{L}^\star\|_2\|X^{(j)\star}\|_{\ell_2}\|\mathsf{E}^{(t)}\|_{\ell_2} + \alpha_1\|X^{(j)\star}\|_2\|\mathsf{E}^{(t)}\|^2_{\ell_2} \leq s\alpha_2/\alpha_0 \leq s\alpha_2\|\mathsf{E}^{(t)}\|_{\ell_2} \leq \sqrt{\Lambda}/20$. Fourth note that the assumptions imply $\Delta/\Lambda \leq 1/6$. Hence by Proposition 2.3.3 applied to $\{X^{(j)\star}\}_{j=1}^n$ and $\{\hat{X}^{(j)}\}_{j=1}^n$ with the choice of $\mathsf{Q} = \mathsf{M}^{(t)}$ we have

$$\mathsf{X}^\star \circ \hat{\mathsf{X}}^+ = (\mathsf{I} + \mathsf{D}) \circ \mathsf{M}^{(t)},$$

where

$$\begin{aligned}\mathsf{D} :=& \frac{1}{n\Lambda}\sum_{j=1}^n ([[(\mathcal{L}^{\star\prime}_{\mathcal{T}^{(j)}}\mathcal{L}^{\star}_{\mathcal{T}^{(j)}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^{\star} \circ \mathsf{E}^{(t)}](X^{(j)\star})) \boxtimes X^{(j)\star} \\ &- \frac{1}{n\Lambda}\sum_{j=1}^n G^{(j)} \boxtimes X^{(j)\star} + \mathsf{F},\end{aligned}$$

and

$$\begin{aligned}\|\mathsf{F}\|_{\ell_2} &\leq 20p(s\alpha_2\|\mathsf{E}^{(t)}\|_{\ell_2})^2/\Lambda + 2p(s\alpha_2\|\mathsf{E}^{(t)}\|_{\ell_2})\Delta/\Lambda^{3/2} \\ &\leq \alpha_5\|\mathsf{E}^{(t)}\|^2_{\ell_2} + \alpha_6(\Delta/\Lambda)\|\mathsf{E}^{(t)}\|_{\ell_2}.\end{aligned} \tag{2.19}$$

[Applying Proposition 2.3.4]: We proceed to bound $\|\mathsf{D}\|_{\ell_2}$. Given a collection $\{A^{(j)}\}_{j=1}^n, \{B^{(j)}\}_{j=1}^n \subset \mathbb{R}^{p\times p}$ one has $\frac{1}{n}\|\sum_{j=1} A^{(j)} \boxtimes B^{(j)}\|_{\ell_2} \leq \max_j \|A^{(j)} \boxtimes B^{(j)}\|_{\ell_2} = \max_j \|A^{(j)}\|_{\ell_2}\|B^{(j)}\|_{\ell_2}$. By combining this inequality with Lemma 2.3.9 we obtain the bounds

$$\begin{aligned}&\frac{1}{n\Lambda}\left\|\sum_{j=1}^n \left(\left[[(\mathcal{L}^{\star\prime}_{\mathcal{T}^{(j)}}\mathcal{L}^{\star}_{\mathcal{T}^{(j)}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^{\star} \circ \mathsf{E}^{(t)}\right](X^{(j)\star})\right) \boxtimes X^{(j)\star}\right\|_{\ell_2} \\ \leq\ & (2s^2 r\|\mathcal{L}^\star\|_2/\Lambda)\|\mathsf{E}^{(t)}\|_{\ell_2} \leq \alpha_3\|\mathsf{E}^{(t)}\|_{\ell_2},\end{aligned} \tag{2.20}$$

and

$$(1/n\Lambda)\left\|\sum_{j=1}^{n} G^{(j)} \boxtimes X^{(j)\star}\right\|_{\ell_2} \leq (\alpha_1 s^2\sqrt{r}/\Lambda)\|\mathsf{E}^{(t)}\|_{\ell_2}^2 \leq \alpha_4\|\mathsf{E}^{(t)}\|_{\ell_2}^2. \tag{2.21}$$

Hence by combining (2.19), (2.20), and (2.21) we have $\|\mathsf{D}\|_{\ell_2} \leq \alpha_3\|\mathsf{E}^{(t)}\|_{\ell_2} + \alpha_4\|\mathsf{E}^{(t)}\|_{\ell_2}^2 + \alpha_5\|\mathsf{E}^{(t)}\|_{\ell_2}^2 + \alpha_6(\Delta/\Lambda)\|\mathsf{E}^{(t)}\|_{\ell_2} \leq \alpha_7\|\mathsf{E}^{(t)}\|_{\ell_2} \leq \alpha_7/\alpha_0 \leq 1/10$. Consequently, by applying Proposition 2.3.4 with this choice of D, we have

$$\mathtt{X}^\star \circ \hat{\mathtt{X}}^+ = (\mathsf{I} + \mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D}) + \mathsf{H}) \circ \mathsf{W} \circ \mathsf{M}^{(t)}, \quad \|\mathsf{H}\|_{\ell_2} \leq (5\alpha_7^2/\sqrt{p})\|\mathsf{E}^{(t)}\|_{\ell_2}^2, \tag{2.22}$$

for some rank-preserver W.

[Conclusion]: Recall from the description of the algorithm that the next iterate is given by $\mathcal{L}^{(t+1)} = \mathcal{L}^\star \circ \mathtt{X}^\star \circ \hat{\mathtt{X}}^+ \circ \mathsf{N}_{\mathcal{L}^\star \circ \mathtt{X}^\star \circ \hat{\mathtt{X}}^+}$, where $\mathsf{N}_{\mathcal{L}^\star \circ \mathtt{X}^\star \circ \hat{\mathtt{X}}^+}$ is the unique positive definite rank-preserver that normalizes $\mathcal{L}^\star \circ \mathtt{X}^\star \circ \hat{\mathtt{X}}^+$. We define $\mathsf{E}^{(t+1)} := \mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D}) + \mathsf{H}$, and hence

$$\mathcal{L}^{(t+1)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t+1)}) \circ \mathsf{M}^{(t+1)}, \tag{2.23}$$

where $\mathsf{M}^{(t+1)} = \mathsf{W} \circ \mathsf{M}^{(t)} \circ \mathsf{N}_{\mathcal{L}^\star \circ \mathtt{X}^\star \circ \hat{\mathtt{X}}^+}$ is a composition of rank-preservers, and hence is also a rank-preserver. It remains to bound $\|\mathsf{E}^{(t+1)}\|_{\ell_2}$.

As $\|[(\mathcal{L}^{\star\prime}_{\mathcal{T}^{(j)}}\mathcal{L}^\star_{\mathcal{T}^{(j)}})^{-1}]_{\mathbb{R}^{p\times p}}\|_2 \leq 2$ from Lemma 2.3.9, we have $[(\mathcal{L}^{\star\prime}_{\mathcal{T}^{(j)}}\mathcal{L}^\star_{\mathcal{T}^{(j)}})^{-1}]_{\mathbb{R}^{p\times p}} \leq 2\mathcal{P}_{\mathcal{T}^{(j)}}$, and hence $(X^{(j)\star}\boxtimes X^{(j)\star}) \otimes [(\mathcal{L}^{\star\prime}_{\mathcal{T}^{(j)}}\mathcal{L}^\star_{\mathcal{T}^{(j)}})^{-1}]_{\mathbb{R}^{p\times p}} \leq 2(X^{(j)\star}\boxtimes X^{(j)\star}) \otimes \mathcal{P}_{\mathcal{T}^{(j)}}$. Moreover, since $(X^{(j)\star}\boxtimes X^{(j)\star}) \otimes [(\mathcal{L}^{\star\prime}_{\mathcal{T}^{(j)}}\mathcal{L}^\star_{\mathcal{T}^{(j)}})^{-1}]_{\mathbb{R}^{p\times p}}$ and $2(X^{(j)\star}\boxtimes X^{(j)\star}) \otimes \mathcal{P}_{\mathcal{T}^{(j)}}$ are Kronecker products of positive semidefinite operators, they too are positive semidefinite operators, and hence $\mathcal{P}_{\mathcal{W}^\perp} \circ (\frac{1}{n}\sum_{j=1}^n (X^{(j)\star} \boxtimes X^{(j)\star}) \otimes [(\mathcal{L}^{\star\prime}_{\mathcal{T}^{(j)}}\mathcal{L}^\star_{\mathcal{T}^{(j)}})^{-1}]_{\mathbb{R}^{p\times p}})^2 \circ \mathcal{P}_{\mathcal{W}^\perp} \preceq \mathcal{P}_{\mathcal{W}^\perp} \circ (\frac{2}{n}\sum_{j=1}^n (X^{(j)\star} \boxtimes X^{(j)\star}) \otimes \mathcal{P}_{\mathcal{T}^{(j)}})^2 \circ \mathcal{P}_{\mathcal{W}^\perp}$. This implies the bound

$$2\Omega \geq \left\| \mathcal{P}_{\mathcal{W}^\perp} \circ \left(\frac{1}{n}\sum_{j=1}^{n} (X^{(j)\star} \boxtimes X^{(j)\star}) \otimes [(\mathcal{L}^{\star\prime}_{\mathcal{T}^{(j)}}\mathcal{L}^\star_{\mathcal{T}^{(j)}})^{-1}]_{\mathbb{R}^{p\times p}} \right) \right\|_2 .$$

Combining this bound with the identity $\mathsf{L}(X_1) \boxtimes X_2 = \mathsf{L} \circ (X_1 \boxtimes X_2)$ we obtain

$$\begin{aligned}
& \frac{1}{n\Lambda}\left\| \mathcal{P}_{\mathcal{W}^\perp}\left(\sum_{j=1}^{n} \left(\left[[(\mathcal{L}^{\star\prime}_{\mathcal{T}^{(j)}}\mathcal{L}^\star_{\mathcal{T}^{(j)}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}^{(t)} \right] (X^{(j)\star}) \right) \boxtimes X^{(j)\star} \right) \right\|_{\ell_2} \\
= \; & \frac{1}{n\Lambda}\left\| \left[\mathcal{P}_{\mathcal{W}^\perp} \circ \left(\sum_{j=1}^{n} (X^{(j)\star} \boxtimes X^{(j)\star}) \otimes [(\mathcal{L}^{\star\prime}_{\mathcal{T}^{(j)}}\mathcal{L}^\star_{\mathcal{T}^{(j)}})^{-1}]_{\mathbb{R}^{p\times p}} \right) \right] (\mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}^{(t)}) \right\|_{\ell_2} \\
\leq \; & (2\Omega/\Lambda)\|\mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}^{(t)}\|_{\ell_2} \leq (2\Omega/\Lambda)\|\mathcal{L}^\star\|_2^2\|\mathsf{E}^{(t)}\|_{\ell_2}.
\end{aligned} \tag{2.24}$$

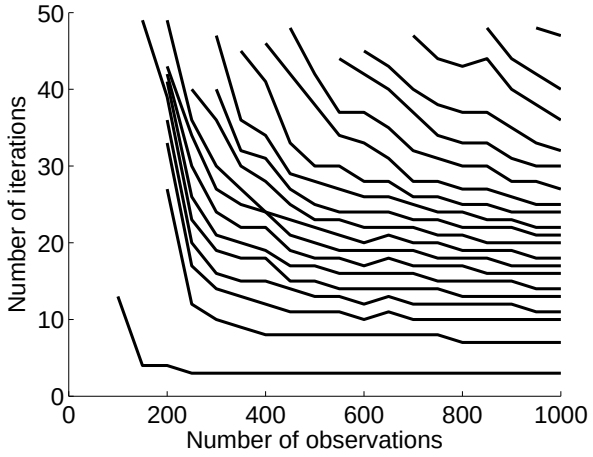


Figure 2.1: Average number of iterations required to identify correct regularizer as a function of the number of observations; each line represents a fixed noise level σ denoting the amount of corruption in the initial guess (see Section 2.4.1 for details of the experimental setup).

From the definition of $\mathsf{E}^{(t+1)}$ we have the relation

$$\begin{aligned}\mathsf{E}^{(t+1)} \;=\; \mathcal{P}_{\mathcal{W}^{\perp}}\Bigg(\frac{1}{n\Lambda}\sum_{j=1}^{n}[[(\mathcal{L}^{\star\prime}_{\mathcal{T}^{(j)}}\mathcal{L}^{\star}_{\mathcal{T}^{(j)}})^{-1}]_{\mathbb{R}^{p\times p}}\circ\mathcal{L}^{\star\prime}\mathcal{L}^{\star}\circ\mathsf{E}^{(t)}]\big(X^{(j)\star}\big)\boxtimes X^{(j)\star} \\ + \frac{1}{n\Lambda}\sum_{j=1}^{n} G^{(j)}\boxtimes X^{(j)\star} + \mathsf{F}\Bigg) + \mathsf{H}. \qquad (2.25)\end{aligned}$$

Since $\mathcal{P}_{\mathcal{W}^{\perp}}$ defines a projection, we have $(1/n\Lambda)\|\mathcal{P}_{\mathcal{W}^{\perp}}(\sum_{j=1}^{n} G^{(j)}\boxtimes X^{(j)\star})\|_{\ell_2} \leq (1/n\Lambda)\|\sum_{j=1}^{n} G^{(j)}\boxtimes X^{(j)\star}\|_{\ell_2}$, and $\|\mathcal{P}_{\mathcal{W}^{\perp}}(\mathsf{F})\|_{\ell_2} \leq \|\mathsf{F}\|_{\ell_2}$. Hence, by applying the bounds (2.19), (2.21), (2.22), and (2.24) to (2.25), we obtain

$$\begin{aligned}\|\mathsf{E}^{(t+1)}\|_{\ell_2} &\leq ((2\Omega/\Lambda)\|\mathcal{L}^{\star}\|_2^2 + \alpha_6(\Delta/\Lambda))\|\mathsf{E}^{(t)}\|_{\ell_2} + (\alpha_4+\alpha_5+5\alpha_7^2/\sqrt{p})\|\mathsf{E}^{(t)}\|_{\ell_2}^2 \\ &= \gamma_0\|\mathsf{E}^{(t)}\|_{\ell_2} + \gamma_1\|\mathsf{E}^{(t)}\|_{\ell_2}^2.\end{aligned}$$

This completes the proof. □

2.4 Numerical Experiments

2.4.1 Illustration with Synthetic Data

We begin with a demonstration of the utility of our algorithm in recovering a regularizer from synthetic data. Our experiment qualitatively confirms the predictions of Theorem 2.3.5 regarding the rate of convergence.

Setup. We generate a standard Gaussian linear map $\mathcal{L} : \mathbb{R}^{7\times 7} \to \mathbb{R}^{30}$ and we normalize it; denote the normalized version as $\mathcal{L}^{\star}$. We generate data $\{\mathbf{y}^{(j)}\}_{j=1}^{1000}$

as $\mathbf{y}^{(j)} = \mathcal{L}^\star(\mathbf{u}^{(j)}\mathbf{v}^{(j)\prime})$, where each $\mathbf{u}^{(j)}, \mathbf{v}^{(j)}$ is drawn independently from the Haar measure on the unit sphere in $\mathbb{R}^7$. We generate standard Gaussian maps $\mathcal{E}^{(i)} : \mathbb{R}^{7\times 7} \to \mathbb{R}^{30}$, $i = 1, \ldots, 20$ that are used to corrupt $\mathcal{L}^\star$ in providing the initial guess to our algorithm. Specifically, for each $\sigma \in \{0.125, 0.25, \ldots, 2.5\}$ and each $\mathcal{E}^{(i)}$, $i = 1, \ldots, 20$ we supply as initial guess to our algorithm the normalized version of $\mathcal{L}^\star + \sigma\mathcal{E}^{(i)}$. In addition we supply the subset $\{\mathbf{y}^{(j)}\}_{j=1}^m$ for each $m \in \{50, 100, \ldots, 1000\}$ to our algorithm. The objective of this experiment is to investigate the role of the number of data points (denoted by m) and the size of the error in the initial guess (denoted by σ) on the performance of our algorithm.

Characterizing recovery of correct regularizer. Before discussing the results, we describe a technique assessing whether our algorithm recovers the correct regularizer. In particular, as we do not know of a tractable technique for computing the distance measure ξ between two linear maps (2.8), we consider an alternative approach for computing the 'distance' between two linear maps. For linear maps from $\mathbb{R}^{p\times p}$ to $\mathbb{R}^d$, we fix a set of unit-Euclidean-norm rank-one matrices $\{\mathbf{s}^{(k)}\mathbf{t}^{(k)\prime}\}_{k=1}^\ell$, where each $\mathbf{s}^{(k)}, \mathbf{t}^{(k)} \in \mathbb{R}^p$ is drawn uniformly from the Haar measure on the sphere and ℓ is chosen to be larger than p^2. Given an estimate $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ of a linear map $\mathcal{L}^\star : \mathbb{R}^{p\times p} \to \mathbb{R}^d$, we compute the following

$$\operatorname{dist}_{\mathcal{L}^\star}(\mathcal{L}) := \frac{1}{\ell}\sum_{k=1}^{\ell} \inf_{\substack{X\in\mathbb{R}^{p\times p} \\ \operatorname{rank}(X)\leq 1}} \left\| \mathcal{L}^\star\left(\mathbf{s}^{(k)}\mathbf{t}^{(k)\prime}\right) - \mathcal{L}(X) \right\|_{\ell_2}^2 . \tag{2.26}$$

To compute the minimum for each term in the sum, we employ the heuristic described in Algorithm 2. If $\mathcal{L}^\star$ satisfies a suitable restricted isometry condition for rank-one matrices and if $\mathcal{L}$ is specified as $\mathcal{L}^\star$ composed with a near-orthogonal rank-preserver, then we have that $\operatorname{dist}_{\mathcal{L}^\star}(\mathcal{L}) \approx 0$; in the opposite direction, as $\ell > p^2$, we have that $\operatorname{dist}_{\mathcal{L}^\star}(\mathcal{L}) \approx 0$ implies $\xi_{\mathcal{L}^\star}(\mathcal{L}) \approx 0$. In our setting with $p = 7$ we set $\ell = 100$. If our algorithm provides an estimate $\mathcal{L}$ such that $\operatorname{dist}_{\mathcal{L}^\star}(\mathcal{L}) < 10^{-3}$, then we declare that our method has succeeded in recovering the correct regularizer.

Results. In Figure 2.1 we plot for each $\sigma \in \{0.125, 0.25, \ldots, 2.5\}$ the average number of iterations – taken over the 20 different initial guesses specified by the normalized versions of $\mathcal{L}^\star + \sigma\mathcal{E}^{(i)}$, $i = 1, \ldots, 20$ – required for Algorithm 3 (with Step 1 computed by solving (2.5) via Algorithm 2) to succeed in recovering the correct regularizer as a function of the number of data points m supplied as input. The different curves in the figure correspond to different noise levels (specified by σ) in the initial guess; that is, the curves higher up in the figure are associated to larger

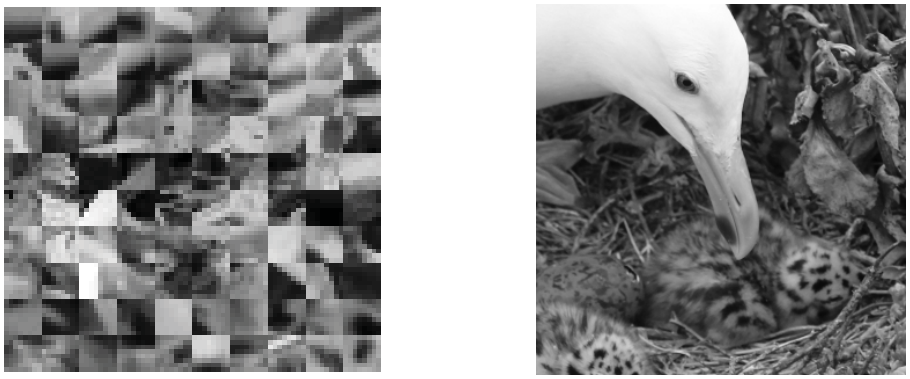


Figure 2.2: Image patches (left) obtained from larger raw images (sample on the right).

noise levels. There are two main conclusions to be drawn from this result. First, the average number of iterations grows as the initial guess is of increasingly poorer quality. Second, and more interesting, is that the number of iterations required for convergence improves with an increase in the number of input data points, but only up to a certain stage beyond which the convergence rate seems to plateau (this is a feature at every noise level in this plot). This observation confirms the predictions of Theorem 2.3.5 and of Proposition 2.3.8 (specifically, see the discussion immediately following this proposition).

2.4.2 Illustration with Natural Images

2.4.2.1 Representing Natural Image Patches

The first stage of this experiment contrasts projections of low-rank matrices and projections of sparse vectors purely from the perspective of representing a collection of image patches.

Setup. We consider a data set $\{\mathbf{y}^{(j)}\}_{j=1}^{6480} \in \mathbb{R}^{64}$ of image patches. This data is obtained by taking 8×8 patches from larger images of seagulls and considering these patches as well as their rotations, as is common in the dictionary learning literature; Figure 2.2 gives an example of a seagull image as well as several smaller patches. To ensure that we learned a centered and suitably isotropic norm, we center the entire data set to ensure that the average of the $\mathbf{y}^{(j)}$'s is the origin and then scale each datapoint so that it has unit Euclidean norm. We apply Algorithm 3 (with Step 1 computed by solving (2.5) via Algorithm 2) and the analog of this procedure for dictionary learning described in Section 2.2.4. We assess the quality of the description of the data set $\{\mathbf{y}^{(j)}\}_{j=1}^{6480}$ as a projection of low-matrices (obtained

using our approach) as opposed to a projection of sparse vectors (obtained using dictionary learning).

Representation complexity. To assess the performance of each representation framework, we require a characterization of the number of parameters needed to specify an image patch in each representation as well as the resulting quality of approximation. Given a collection $\{\mathbf{y}^{(j)}\}_{j=1}^{n} \subset \mathbb{R}^d$, suppose we represent each point as $\mathbf{y}^{(j)} \approx \mathcal{L}(X^{(j)})$ for a linear map $\mathcal{L} : \mathbb{R}^{p\times p} \to \mathbb{R}^d$ and a rank-r matrix $X^{(j)} \in \mathbb{R}^{p\times p}$. The number of parameters required to specify each $X^{(j)}$ is $2pr - r^2$ and the number of parameters required to specify $\mathcal{L}$ is dp^2. Consequently, the average number of parameters required to specify each $\mathbf{y}^{(j)}$ is $2pr - r^2 + \frac{dp^2}{n}$. In a similar manner, if each $\mathbf{y}^{(j)} \approx L\mathbf{x}^{(j)}$ for a linear map $L : \mathbb{R}^p \times \mathbb{R}^d$ and a vector $\mathbf{x}^{(j)} \in \mathbb{R}^p$ with s nonzero coordinates, the average number of parameters required to each $\mathbf{y}^{(j)}$ is $2s + \frac{dp}{n}$. In each case, we assess the quality of the approximation by considering the average squared error over the entire set $\{\mathbf{y}^{(j)}\}_{j=1}^{n}$.

Results. We initialize both our algorithm and the dictionary learning method with random linear maps (suitably normalized in each case). Before contrasting the two approaches we highlight the improvement in performance our method provides over a pure random linear map. Specifically, Figure 2.3 shows for several random initializations that our algorithm (as well as the alternating update method in dictionary learning) provides a significant refinement in approximation quality as the number of iterations increases. Therefore, there is certainly value in employing our algorithm (even with a random initialization) to obtain better representations than pure random projections of low-rank matrices. Next we proceed to a detailed comparison of the two representation frameworks. We employ our approach to learn a representation of the image patch data set with $p \in \{9, 10, \ldots, 15\}$ and the values of the rank r chosen so that the overall representation complexity lies in the range $[17, 33]$. Similarly, we employ dictionary learning with $p \in \{100, 200, \ldots, 1400\}$ and the values of the sparsity level s chosen so that the overall representation complexity lies in the range $[17, 33]$. The left subplot in Figure 2.5 gives a comparison of these two frameworks. (To interpret the y-axis of the plot, note that the each data point is scaled to have unit norm.) Our approach provides an improvement over dictionary learning for small levels of representation complexity and is comparable at larger levels.

Comparison of atoms. Figure 2.4 gives an illustration of the atoms obtained from classical dictionary learning (i.e., learning a polyhedral regularizer) as well as those learned using our approach. The left subplot shows the finite collection of atoms of

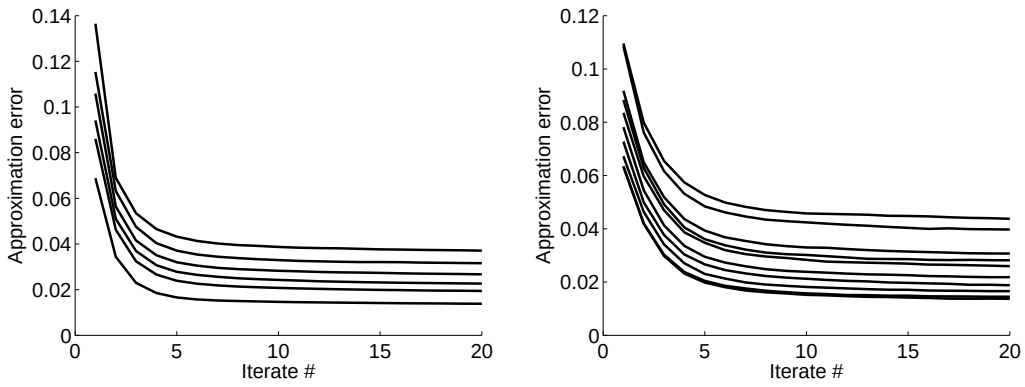


Figure 2.3: Progression in mean-squared error with increasing number of iterations with random initializations for learning a semidefinite regularizer (left) and a polyhedral regularizer (right).

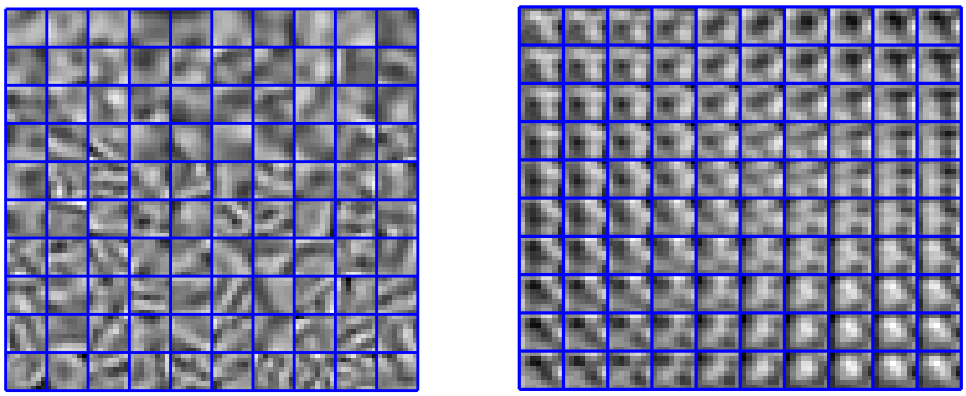


Figure 2.4: Comparison between atoms learned from dictionary learning (left) and our algorithm (right).

a polyhedral regularizer (corresponding to the finite number of extreme points), and the right subplot shows a finite subset of the infinite collection of atoms learned using our approach. The individual atoms in each case generally correspond to piecewise smooth regions separated by boundaries. However, the geometry of the *collection* of atoms is distinctly different in the two cases; in particular, the atoms learned using our approach better represent the transformations underlying natural images. As we discuss in the next set of experiments, our framework provides regularizers that lead to improved denoising performance on natural images in comparison with polyhedral regularizers.

2.4.2.2 Denoising Natural Image Patches

We compare the performance of polyhedral and semidefinite regularizers in denoising natural image patches corrupted by noise.

Setup. The 6480 data points from the previous experiment are designated as a training set. Here we consider an additional collection $\{\mathbf{y}_{\text{test}}^{(j)}\}_{j=1}^{720} \subset \mathbb{R}^{64}$ of 8×8 test image patches obtained from larger seagull images (as with the training set), and subsequently shifted by an average of the pre-centered training set. We corrupt each of these test points by i.i.d. Gaussian noise to obtain $\mathbf{y}_{\text{obs}}^{(j)} = \mathbf{y}_{\text{test}}^{(j)} + \boldsymbol{\varepsilon}^{(j)}$, $j = 1, \ldots, 720$, where each $\boldsymbol{\varepsilon}^{(j)} \sim \mathcal{N}(0, \sigma^2 I)$ with σ^2 chosen so that the average signal-to-noise ratio $\frac{1}{720}\sum_{j=1}^{n} \frac{\|\mathbf{y}_{\text{test}}^{(j)}\|_{\ell_2}^2}{64\sigma^2} \approx 18$. Our objective is to investigate the denoising performance of the polyhedral and semidefinite regularizers (learned on the training set) on the data set $\{\mathbf{y}_{\text{obs}}^{(j)}\}_{j=1}^{720}$. Specifically, we analyze the following proximal denoising procedure:

$$\hat{\mathbf{y}}_{\text{denoise}} = \underset{\mathbf{y} \in \mathbb{R}^{64}}{\arg\min} \quad \tfrac{1}{2}\|\mathbf{y}_{\text{obs}} - \mathbf{y}\|_{\ell_2}^2 + \lambda \|\mathbf{y}\|, \tag{2.27}$$

where $\|\cdot\|$ is a regularizer learned on the training set and $\lambda > 0$ is a regularization parameter.

Computational complexity of regularizer. To compare the performances of different regularizers, it is instructive to consider the cost associated with employing a regularizer for denoising. In particular, the regularizers learned on the training set have unit-balls that are specified as linear images of the nuclear norm ball and the ℓ_1 ball. Consequently, the main cost associated with employing a regularizer is the computational complexity of solving the corresponding proximal denoising problem (5.5). Thus, we analyze the normalized mean-squared denoising error $\frac{1}{720}\sum_{j=1}^{n} \frac{\|\mathbf{y}_{\text{obs}}^{(j)} - \mathbf{y}_{\text{denoise}}^{(j)}\|_{\ell_2}^2}{64\sigma^2}$ of a regularizer as a function of the computational complexity of solving (5.5). For a polyhedral norm $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$ with unit ball specified as the image under a linear map $L : \mathbb{R}^q \to \mathbb{R}^d$ of the ℓ_1 ball in $\mathbb{R}^q$, we solve (5.5) as follows by representing the norm $\|\cdot\|$ in a lifted manner:

$$\begin{aligned} \hat{\mathbf{y}}_{\text{denoise}} = \; & \underset{\substack{\mathbf{x},\mathbf{z} \in \mathbb{R}^q \\ s,t \in \mathbb{R}}}{\arg\min} \quad \tfrac{1}{2}s + \lambda t \\ & \text{s.t.} \quad \|\mathbf{y}_{\text{obs}} - L\mathbf{x}\|_{\ell_2}^2 \le s, \quad \sum_{i=1}^{p} \mathbf{z}_i \le t, \quad \begin{pmatrix} \mathbf{z} - \mathbf{x} \\ \mathbf{z} + \mathbf{x} \end{pmatrix} \ge 0. \end{aligned} \tag{2.28}$$

To solve (2.28) to an accuracy ϵ using an interior-point method with the usual logarithmic barriers for the nonnegative orthant and the second-order cone, we have

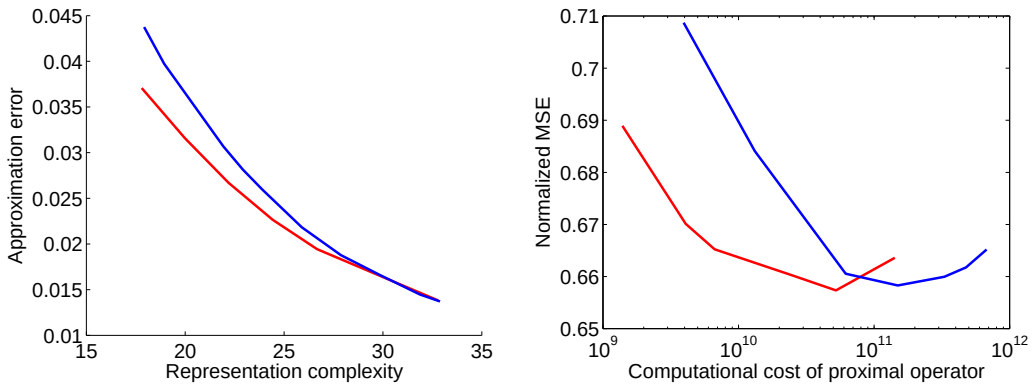


Figure 2.5: Comparison between dictionary learning (blue) and our approach (red) in representing natural image patches (left); comparison between polyhedral (blue) and semidefinite (right) regularizers in denoising natural image patches (right).

that the number of operations required is $\sqrt{2q+2}\log(\frac{2q+2}{\epsilon\eta}((d+2q+2)^3+(2q+2)^3))$ – this represents the number of outer loop iterations of the interior point method – multiplied by $(d+2q+2)^3+(2q+2)^3$ – this represents the number of operations required to solve the associated linear system in the inner loop – for a barrier parameter η [103, 118]. In a similar manner, for a semidefinite regularizer $\|\cdot\|:\mathbb{R}^d\to\mathbb{R}$ with unit ball specified as the image under a linear map $\mathcal{L}:\mathbb{R}^{p\times p}\to\mathbb{R}^d$ of the nuclear norm ball in $\mathbb{R}^{p\times p}$, we again solve (5.5) as follows by representing the norm $\|\cdot\|$ in an analogous lifted manner:

$$\begin{aligned}\hat{\mathbf{y}}_{\text{denoise}} = \underset{\substack{X\in\mathbb{R}^{p\times p}\\ Z_1,Z_2\in\mathbb{S}^p\\ s,t\in\mathbb{R}}}{\arg\min}\ \ &\tfrac{1}{2}s+\lambda t\\ \text{s.t.}\ \ &\|\mathbf{y}_{\text{obs}}-\mathcal{L}(X)\|_{\ell_2}^2\le s,\ \tfrac{1}{2}\operatorname{tr}(Z_1+Z_2)\le t,\ \begin{pmatrix}Z_1 & X\\ X' & Z_2\end{pmatrix}\succeq 0.\end{aligned} \tag{2.29}$$

As before, to solve (2.29) to an accuracy ϵ using an interior-point method with the usual logarithmic barriers for the positive-semidefinite cone and the second-order cone, we have that the number of operations required is $\sqrt{2p+2}\log(\frac{2p+2}{\epsilon\eta}((d+2\binom{p}{2}+2)^3+(2\binom{p}{2}+2)^3))$ multiplied by $(d+\binom{2p}{2}+2)^3+(\binom{2p}{2}+2)^3$ for a barrier parameter η [118].

Results. We learn semidefinite regularizers on the training set using Algorithm 3 for $p\in\{9,\dots,20\}$ and for a rank of 1. We also learn polyhedral regularizers on the training set using dictionary learning for $q\in\{9^2,10^2,\dots,20^2\}$ and with corresponding sparsity levels in the range $\{\sqrt{q}-1,\sqrt{q}\}$ to ensure that the representation

complexity matches the corresponding representation complexity of the images of rank-one matrices in the semidefinite case. As the lifted dimensions p^2 and q increase, the computational complexities of the associated proximal denoisers (with the learned regularizers) also increase. The right subplot in Figure 2.5 gives the average normalized mean-squared error over the noisy test data (generated as described above). The optimal choice of the regularization parameter λ for each regularizer is obtained by sweeping over a range to obtain the best denoising performance, as we have access to the underlying uncorrupted image patches $\{\mathbf{y}_{\text{test}}^{(j)}\}_{j=1}^{720}$. For both types of regularizers the denoising performance improves initially before degrading due to overfitting. More significantly, given a fixed computational budget, these experiments suggest that semidefinite regularizers provide better performance than polyhedral regularizers in denoising image patches in our data set. The denoising operation (5.5) is in fact a basic computational building block (often referred to as a proximal operator) in first-order algorithms for solving convex programs that arise in a range of inverse problems [109]. As such, we expect the results of this section to be qualitatively indicative of the utility of our approach in other inferential tasks beyond denoising.

2.5 Discussion

Our paper describes an algorithmic framework for learning regularizers from data in settings in which prior domain-specific expertise is not directly available. We learn these regularizers by computing a structured factorization of the data matrix, which is accomplished by combining techniques for the affine rank minimization problem with the Operator Sinkhorn scaling procedure. The regularizers obtained using our method are convex, and they can be computed via semidefinite programming. Our approach may be viewed as a semidefinite analog of dictionary learning, which can be interpreted as a technique for learning polyhedral regularizers from data. We discuss next some directions for future work.

2.5.1 Algorithmic questions

It would be of interest to better understand the question of initialization for our algorithm. Random initialization often works well in practice and it would be useful to provide theoretical support for this approach by building on recent work on other factorization problems [64, 141]. To this end, we describe two experimental setups on synthetic data showing instances where our algorithm recovers the true regularizer from random initialization. In the first setup we generate a standard Gaussian linear

map $\mathcal{L} : \mathbb{R}^{8\times 8} \to \mathbb{R}^{50}$ and normalize it. Let $\mathcal{L}^\star$ denote the resulting normalized map. We generate data $\{\mathbf{y}^{(j)}\}_{j=1}^{10^4}$ as $\mathbf{y}^{(j)} = \mathcal{L}^\star(\mathbf{u}^{(j)}\mathbf{v}^{(j)\prime})/\|\mathcal{L}^\star(\mathbf{u}^{(j)}\mathbf{v}^{(j)\prime})\|_{\ell_2}$, where each $\mathbf{u}^{(j)}, \mathbf{v}^{(j)}$ is drawn independently from the Haar measure on the unit sphere in $\mathbb{R}^8$. We apply our algorithm to the data, and we supply as initialization the normalization of a standard Gaussian linear map. The left subplot of Figure 2.6 shows the progression of the mean-squared error over 10 different initializations. As the measurements do not contain any additional noise, the minimum attainable error is zero. We observe that our algorithm recovers the regularizer in all 10 random initializations; moreover, we observe local, linear convergence in the neighborhood of the global minimizer, which agrees with our analysis. Note that the progress of our algorithm reveals interesting behavior in that the global recovery of the regularizer is characterized by three distinct phases – (*i*) an initial phase in which progress is significant; (*ii*) an intermediate phase in which progress is incremental but stable; and (iii) a terminal phase that corresponds to local, linear convergence. In particular, these graphs indicate that global convergence to the underlying regularizer is *not* linear. The second setup is similar to the first one, with the two main differences being that we consider a linear map $\mathcal{L}^\star : \mathbb{R}^{8\times 8} \to \mathbb{R}^{60}$ of slightly different dimensions, and that the data points $\{\mathbf{y}^{(j)}\}_{j=1}^{2\times 10^4}$ are images of rank-two matrices. The right subplot of Figure 2.6 shows the progression of our algorithm over 10 different initializations. In contrast to the previous setup where every initialization led to a global minimum, in this case our algorithm attains a local minimum in 4 out of 10 initializations and a global minimum in the remaining 6 initializations. In summary, our experiments suggest that random initialization may sometimes be effective, and understanding this effectiveness warrants further investigation.

Beyond random initialization, there have also been efforts on data-driven strategies for initialization in dictionary learning by reducing the question to a type of clustering / community detection problem [3, 8]. While the relation between clustering and estimating the elements of a finite atomic set is conceptually natural, identifying an analog of the clustering problem for estimating the image of a variety of rank-one matrices (which is a structured but infinite atomic set) is less clear; we seek such a conceptual link in order to develop an initialization strategy for our algorithm. In a completely different direction, there is also recent work on a convex relaxation for the dictionary learning problem that avoids the difficulties associated with local minima [11]; while this technique is considerably more expensive computationally in comparison with alternating updates, developing analogous convex relaxation approaches for the problem of learning semidefinite regularizers may subsequently

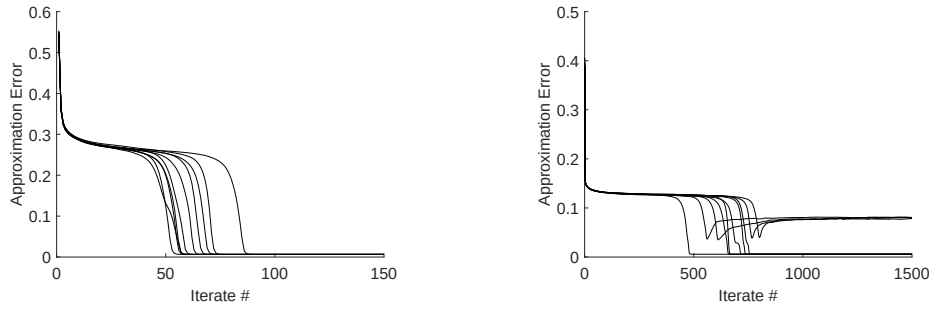


Figure 2.6: Progression of our algorithm in recovering regularizers in a synthetic experimental set-up; the horizontal axis represents the number of iterations, and each line corresponds to a different random initialization. The left subplot shows a problem instance in which all 10 different random initializations recover a global minimizer, while the right subplot shows a different problem instance in which 4 out of 10 random initializations lead to local minima.

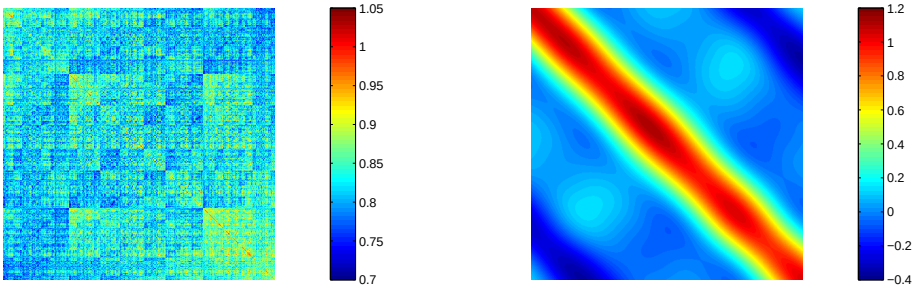


Figure 2.7: Gram matrices of images of sparse vectors (left) and low-rank matrices (right).

point the way to efficient global techniques that are different from alternating updates.

2.5.2 Approximation-theoretic questions

The focus of our paper has been on the algorithmic aspects of learning semidefinite regularizers from data. It is of interest to investigate the power of finite atomic sets in comparison with atomic sets specified as projections of determinantal varieties from a harmonic analysis perspective (for a fixed representation complexity; see Section 2.4.2.1 for a discussion on how these are defined). For example, what types of data are better described using one representation framework versus the other?

Comparison of Gram matrices. As a simple preliminary illustration, we generate two sets of 400 points in $\mathbb{R}^{500}$, with the first set being a random projection of sparse

vectors in $\mathbb{R}^{900}$ and the second set being a random projection of rank-one matrices in $\mathbb{R}^{900}$ of the form $(\cdots \cos(2\pi\alpha_j t_i),\ \sin(2\pi\alpha_j t_i),\ \cdots)'\,(\cdots \cos(2\pi\beta_j t_i),\ \sin(2\pi\beta_j t_i),\ \cdots)$ for randomly chosen frequencies α_j, β_j; the representation complexities of both these sets is the same. Figure 2.7 gives the Gram matrices associated with these data sets. The data set of projections of sparse vectors appears to consist of 'clusters' of 'block' structure, while the data set of projections of low-rank matrices appears to consist of smoother 'toroidal' structure. We seek a better understanding of this phenomenon by analyzing the relative strengths of representations based on finite atomic sets versus projections of low-rank matrices.

Representing group invariant datasets. A more targeted approach for answering our approximation theoretic question is the following: are datasets that possess natural invariances arising from group transformations better described as determinantal varieties compared to finite atomic sets?

As a simple illustration, we consider the following dataset comprising 1000 image patches of dimensions 21×21 generated by rotating a single image at regular intervals. We subsequently project the data to its top three principal components. Figure 2.8 shows a subset of these images.

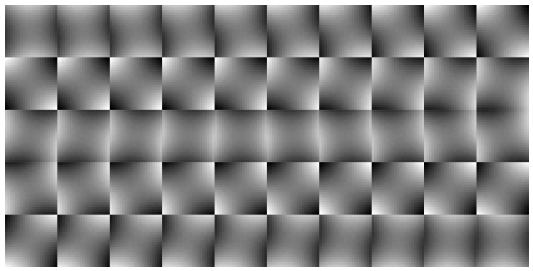


Figure 2.8: Dataset of rotated image patches.

In the first instance, we apply classical dictionary learning to learn a collection of six atoms $\{\pm L_1, \pm L_2, \pm L_3\}$, and we show these in Figure 2.9. Notice that the atoms can be naturally interpreted as rotations of the same image, and whose orientations are spaced across regular intervals.

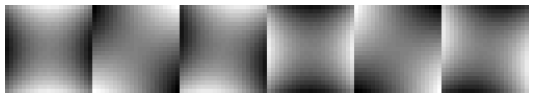


Figure 2.9: A collection of six atoms learned from the data using dictionary learning.

In the second instance, we apply our method to learn an infinite collection of atoms specified as images of rank-one matrices $\{\pm\mathcal{L}(\mathbf{e}\mathbf{e}') : \mathbf{e} \in \mathbb{R}^2, \|\mathbf{e}\|_{\ell_2} = 1\}$,

$\mathcal{L} : \mathbb{S}^2 \to \mathbb{R}^3$, which we show in Figure 2.10 (we omit negations). Notice that the collection of atoms in Figure 2.10 correctly reflects the higher order relationship that is present in the data.

It is perhaps worth noting that the original dataset is *not* synthetically generated to be exactly representable as images of rank-one matrices – Figure 2.11 shows the residual error of expressing the data as projections of rank-one matrices.

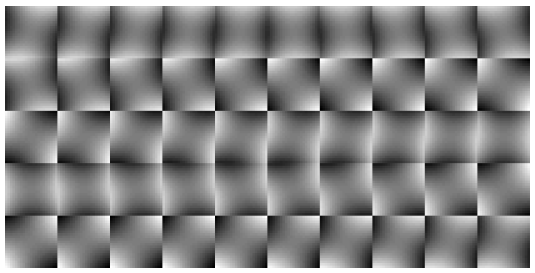


Figure 2.10: A semidefinite representable dictionary learned from the data.

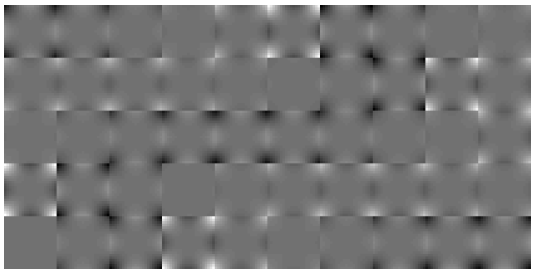


Figure 2.11: Residual error from representing dataset as projections of a rank-one matrix.

General atomic sets. More generally, it is also of interest to explore other families of infinite atomic sets that yield tractable regularizers in other conic programming frameworks. Specifically, dictionary learning and our approach provide linear and semidefinite programming regularizers, but there are other families of computationally efficient convex cones such as the power cone and the exponential cone; learning atomic sets that are amenable to optimization in these frameworks would lead to a broader suite of data-driven approaches for identifying regularizers.

Chapter 3

FITTING TRACTABLE CONVEX SETS TO SUPPORT FUNCTION EVALUATIONS

3.1 Introduction

We consider the problem of estimating a compact convex set given (possibly noisy) evaluations of its support function. Formally, let $\mathcal{K}^\star \subset \mathbb{R}^d$ be a set that is compact and convex. The support function $h_{\mathcal{K}^\star}(\mathbf{u})$ of the set $\mathcal{K}^\star$ evaluated in the direction $\mathbf{u} \in \mathcal{S}^{d-1}$ is defined as:

$$h_{\mathcal{K}^\star}(\mathbf{u}) := \sup_{\mathbf{g} \in \mathcal{K}^\star} \langle \mathbf{g}, \mathbf{u} \rangle.$$

Here $\mathcal{S}^{d-1} := \{\mathbf{g} : \|\mathbf{g}\|_2 = 1\} \subset \mathbb{R}^d$ denotes the $(d-1)$-dimensional unit sphere. In words, the quantity $h_{\mathcal{K}^\star}(\mathbf{u})$ measures the maximum displacement of the plane normal to $\mathbf{u}$ intersecting $\mathcal{K}^\star$. Given a collection of noisy support function evaluations $\left\{(\mathbf{u}^{(i)}, y^{(i)}) : y^{(i)} = h_{\mathcal{K}^\star}(\mathbf{u}^{(i)}) + \varepsilon^{(i)}\right\}_{i=1}^n$, where each $\varepsilon^{(i)}$ denotes additive noise, our goal is to reconstruct a convex set $\hat{\mathcal{K}}$ that is close to $\mathcal{K}^\star$.

The problem of estimating a convex set from support function evaluations arises in a wide range of problems such as computed tomography [115], target reconstruction from laser-radar measurements [91], and projection magnetic resonance imaging [71]. For example, in tomography the extent of the absorption of parallel rays projected onto an object provides support information [115, 135], while in robotics applications support information can be obtained from an arm clamping onto an object in different orientations [115]. A natural approach to fit a compact convex set to support function data is the following least squares estimate (LSE):

$$\hat{\mathcal{K}}_{\text{LSE}} \in \operatorname*{argmin}_{\mathcal{K} \subset \mathbb{R}^d : \mathcal{K} \text{ is compact, convex}} \frac{1}{n} \sum_{i=1}^n \left(y^{(i)} - h_{\mathcal{K}}(\mathbf{u}^{(i)}) \right)^2. \tag{3.1}$$

An LSE always exists and it is not defined uniquely, although it is always possible to select a polytope that is an LSE; this is the choice that is most commonly employed and analyzed in prior work. For example, the algorithm proposed by Prince and Willsky [115] for planar convex sets reconstructs a polyhedral LSE described in terms of its facets, while the algorithm proposed by Gardner and Kiderlen [60] for convex sets in any dimension provides a polyhedral LSE reconstruction described in terms of extreme points. Despite the fact that $\hat{\mathcal{K}}_{\text{LSE}}$ is a consistent estimator of $\mathcal{K}^\star$,

it has a number of significant drawbacks. In particular, as the formulation (3.1) does not incorporate any additional structural information about $\mathcal{K}^\star$ beyond convexity, the estimator $\hat{\mathcal{K}}_{\mathrm{LSE}}$ is not well-behaved when the measurements available are noisy or small in number. Further, even when the number of measurements is large, the complexity of the resulting estimate grows with the number of measurements in the absence of any regularization as the regression problem (3.1) is nonparametric (the collection of all compact convex sets in $\mathbb{R}^d$ is not finitely parametrized); consequently, the facial structure of the reconstruction provides little information about the geometry of the underlying set.[1] Figure 3.1 provides an illustration of these points. Finally, if the underlying set $\mathcal{K}^\star$ is not polyhedral, a polyhedral choice for the solution $\hat{\mathcal{K}}_{\mathrm{LSE}}$ (as is the case with much of the literature on this topic) produces poor reconstructions. Indeed, even for intrinsically "simple" convex bodies such as the Euclidean ball, one necessarily requires many vertices or facets in order to obtain accurate polyhedral approximations; see Figure 3.1.

3.1.1 Our Contributions

We describe a framework for *regularizing* for the complexity of the reconstruction in the formulation (3.1). In particular, we consider reconstructions specified as convex sets in $\mathbb{R}^d$ (the ambient space in which $\mathcal{K}^\star$ lies) that are linear images of concisely described convex sets in $\mathbb{R}^q$, with q not being too much larger than d. The *lifting dimension* q serves as a surrogate for the complexity of the reconstruction. Convex sets described in this manner are significant as there exist computationally efficient algorithms for the optimization of linear functionals over such sets [103]. We employ these ideas in a conceptually different context in the setting of the present paper in order to regularize for the reconstruction in (3.1), which addresses many of the drawbacks with the LSE outlined previously.

Formally, we consider the following regularized convex set regression problem:

$$\hat{\mathcal{K}}_n(C) \in \operatorname*{argmin}_{\mathcal{K}:\mathcal{K}=A(C),A\in L(\mathbb{R}^q,\mathbb{R}^d)} \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - h_{\mathcal{K}}(\mathbf{u}^{(i)})\right)^2. \tag{3.2}$$

Here $C \subset \mathbb{R}^q$ is a compact convex set in $\mathbb{R}^q$ and $L(\mathbb{R}^q,\mathbb{R}^d)$ denote the set of linear maps from $\mathbb{R}^q$ to $\mathbb{R}^d$. In words, the above regularized formulation aims to identify a convex set that is expressible as a linear image of the given set C that minimizes

[1] We note that this is the case even though the estimator $\hat{\mathcal{K}}_{\mathrm{LSE}}$ is consistent; in particular, consistency simply refers to the convergence of $\hat{\mathcal{K}}_{\mathrm{LSE}}$ to $\mathcal{K}^\star$ in a 'metric' sense (e.g., Hausdorff distance) and it does not provide any information about the facial structure of $\hat{\mathcal{K}}_{\mathrm{LSE}}$ relative to that of $\mathcal{K}^\star$.

the error. The choices of C that are most commonly employed in our experimental demonstrations are the simplex and the free spectrahedron.

Example. *The* simplex *in* $\mathbb{R}^q$ *is the set:*

$$\Delta^q := \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^q, \mathbf{x} \geq 0, \langle \mathbf{x}, \mathbf{1} \rangle = 1\} \quad \textit{where} \quad \mathbf{1} = (1, \ldots, 1)'.$$

Convex sets that are expressible as projections of Δ^q *are precisely polytopes with at most* q *extreme points.*

Example. *Let* $\mathbb{S}^p \cong \mathbb{R}^{\binom{p+1}{2}}$ *denote the space of* $p \times p$ *real symmetric matrices. The* free spectrahedron $\mathcal{O}^p \subset \mathbb{S}^p$ *(also called the* spectraplex*) is the set:*

$$\mathcal{O}^p := \{X : X \in \mathbb{S}^p, X \succ 0, \langle X, I \rangle = 1\}, \quad \textit{where} \quad I \in \mathbb{S}^p \textit{ is the identity matrix.}$$

The free spectrahedron is a semidefinite analog of the simplex and the lifting dimension in this case is $\binom{p+1}{2}$*. The free spectrahedron is especially useful in situations in which we seek non-polyhedral reconstructions, as can be seen in Figure 3.1.*

Although our theoretical analysis is largely valid for arbitrary compact convex sets C, we generally operate under the premise that the choice of C employed in (3.2) comes from the families $\{\Delta^q\}_{q=1}^{\infty}$ and $\{\mathcal{O}^p\}_{p=1}^{\infty}$. The specific selection of C from these families is governed by the complexity of the reconstruction one seeks, which is typically based on prior information about the underlying convex set. Our results in Section 3.3 on the statistical properties of the estimator (3.2) rely on the availability of such additional knowledge about the complexity of the underlying set. In practice in the absence of such information, cross-validation may be employed to obtain a suitable reconstruction; see Section 3.6.

In Section 3.2 we discuss preliminary aspects of our technical setup such as properties of the set of minimizers of the problem (3.2) as well as a stylized probabilistic model for noisy support function evaluations. These serve as a basis for the subsequent development in the paper. In Section 3.3 we provide the main theoretical guarantees of our approach. In the first result concerning consistency, we show that the sequence of estimates $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^{\infty}$ converges almost surely (as $n \to \infty$) to that linear image of C that is closest to the underlying set $\mathcal{K}^\star$. Under suitable additional conditions, we also characterize certain asymptotic distributional aspects of the sequence $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^{\infty}$; this result is based on a functional central limit theorem, and it requires the computation of appropriate entropy bounds for Vapnik-Chervonenkis (VC) classes of sets that admit semialgebraic descriptions. Our third result describes

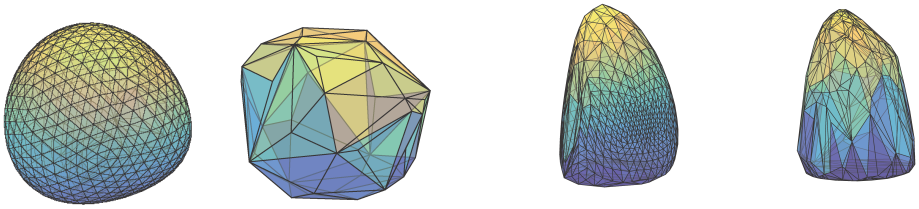
the facial structure of $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^{\infty}$ in relation to the underlying set $\mathcal{K}^\star$. We prove under appropriate conditions that if $\mathcal{K}^\star$ is a polytope then our approach provides a reconstruction that recovers all the simplicial faces (for sufficiently large n); in particular, if $\mathcal{K}^\star$ is a simplicial polytope, we recover a polytope that is combinatorially equivalent to $\mathcal{K}^\star$. Our result is more general, and it applies to non-polyhedral $\mathcal{K}^\star$.

In the sequel, we describe a conceptual connection between the formulation (3.2) (when C is a simplex) and K-means clustering. Accordingly, the algorithms we propose in Section 3.4 for computing $\hat{\mathcal{K}}_n(C)$ – one based on gradient descent, and another based on minimizing in alternating directions – bear significant similarities with prominent methods, such as Lloyd's algorithm, that are widely employed for clustering problems (see Section 3.4). As the problem (3.2) is not convex as formulated, our algorithms are not guaranteed to find a global minimum. Indeed, the connection between (3.2) and K-means clustering suggests that obtaining globally optimal solutions may be NP-hard in general [42, 95]. We discuss this point further in Section 3.6.

A restriction in the development in this paper is that given a lifting dimension q, we do not consider further optimization of the set $C \subset \mathbb{R}^q$ in the formulation (3.2) aside from the choices of the simplex Δ^q and the free spectrahedron $\mathcal{O}^p$ (with $q = \binom{p+1}{2}$). Specifically, both the simplex Δ^q and the free spectrahedron $\mathcal{O}^p$ are particular affine sections of the nonnegative orthant in $\mathbb{R}^q$ and the cone of positive semidefinite matrices in $\mathbb{S}^p$, respectively. Consequently, a natural approach – based on the literature on conic optimization [70, 152] – is to further improve the choice of C by optimizing over all affine slices of the nonnegative orthant in $\mathbb{R}^q$ or the cone of positive semidefinite matrices in $\mathbb{S}^p$. Such an additional degree of flexibility in the formulation (3.2) leads to technical complications with establishing asymptotic normality in Section 3.3.2 as well as to challenges with developing algorithms for solving (3.2) (even to obtain a local optimum). We remark on these difficulties in Section 3.3.2, and for the remainder of the paper we proceed with the framework discussed in the preceding paragraphs.

3.1.2 Estimating Centrally Symmetric Convex Sets

In some applications we may be interested in estimating convex bodies that are known to be *centrally symmetric* – these are convex sets with the property that $-\mathbf{x} \in \mathcal{K}$ whenever $\mathbf{x} \in \mathcal{K}$, and they naturally arise as level sets of *norms*. From a statistical viewpoint, it is frequently beneficial to enforce such a symmetry explicitly



(a) Reconstructions of the unit ℓ_2-ball from 50 noisy support function measurements as the projection of O^3 (our approach, left), and the LSE (right).

(b) Reconstructions of the convex mesh of a human lung from 300 noiseless support function measurements as the projection of O^6 (our approach, left), and the LSE (right).

Figure 3.1: Comparison between our approach and the LSE.

in the estimation procedure, as it reduces the number of degrees of freedom in the problem.

Based on our framework of estimating convex sets as projections of a lifting set $C \subset \mathbb{R}^q$, a natural extension of our ideas to estimating centrally symmetric convex sets $\mathcal{K}$ is to simply choose $C \subset \mathbb{R}^q$ to also be centrally symmetric. The subsequent discussion in this chapter also applies to such a setting. In particular, the natural extension of projections of the simplex are projections of the ℓ_1 ball, and the natural extensions of projections of the free spectrahedron are the projections of the nuclear norm ball over the space of symmetric as well as asymmetric matrices.

3.1.3 Related Work

3.1.3.1 Consistency of Convex Set Regression

There is a well-developed body of prior work on analyzing the consistency of the LSE (3.1). Gardner et al. [61] prove that the (polyhedral) estimates $\hat{\mathcal{K}}_{\text{LSE}}$ converge to the underlying set $\mathcal{K}^\star$ in the Hausdorff metric as $n \to \infty$. A number of related work subsequently analyzed the rate of this convergence in minimax settings [27, 74]. These results hold under relatively minimal assumptions on the available support function evaluations and on the set $\mathcal{K}^\star$ (essentially that this set is convex and compact). In contrast, the consistency result in the present paper corresponding to the constrained estimator (3.2) is qualitatively different. For a given compact and convex set $C \subset \mathbb{R}^q$, we prove that the sequence of estimates $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^{\infty}$ converges to that linear image of C that is closest to the underlying set $\mathcal{K}^\star$; in particular, $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^{\infty}$ only converges to $\mathcal{K}^\star$ if $\mathcal{K}^\star$ can be represented as a linear image of C. However, there are several advantages to the framework presented in this paper in comparison with prior work. First, the constrained estimator (3.2) lends itself

to a precise asymptotic distributional characterization that is unavailable in the unconstrained case (3.1). Second, under appropriate conditions, the constrained estimator (3.2) recovers the facial structure of the underlying set $\mathcal{K}^\star$ unlike $\hat{\mathcal{K}}_{\mathrm{LSE}}$. More significantly, beyond these technical distinctions, the constrained estimator (3.2) also yields concisely described non-polyhedral reconstructions (as well as consistency and asymptotic distributional characterizations) based on linear images of the free spectrahedron, in contrast to the usual choice of a polyhedral LSE in the previous literature.

3.1.3.2 Fitting Convex Sets with Smooth Boundaries

We are aware of a line of work by Fisher et al. [55] on fitting convex sets with smooth boundaries to support function measurements. They propose interpolating between support function evaluations using splines and the von Mises kernel to obtain a smooth estimate of the support function. This smoothing is done in a local fashion and the resulting reconstruction is increasingly complex to describe as the number of measurements available grows. In contrast, our approach to producing non-polyhedral estimates based on fitting linear images of free spectrahedra is more global in nature, and we explicitly regularize for the complexity of our reconstruction based on the choice of the lifting dimension. Further, the reconstructions produced by Fisher et al. [55] are convex sets with smooth boundaries, i.e., the resulting estimates do not possess any faces other than the extreme points and the entire set itself. Consequently, any interesting facial geometry in the underlying set is not captured in these reconstructions, unlike those based on linear images of free spectrahedra as in the present paper.

3.1.3.3 K-means clustering

When the set $C \subset \mathbb{R}^q$ in the constrained estimator (3.2) is the simplex Δ^q, the resulting optimization formulation bears a great deal of similarity to the problem of clustering [93]. Specifically, in K-means clustering, one wishes to partition a collection of points $\{\mathbf{y}^{(i)}\}_{i=1}^n \subset \mathbb{R}^d$ into q disjoint clusters, where the clusters are represented by a collection of 'centers' $\{\mathbf{v}_j\}_{j=1}^q \subset \mathbb{R}^d$. This task is typically posed as the following minimization problem

$$\underset{\{\mathbf{v}_j\}_{j=1}^q \subset \mathbb{R}^d}{\operatorname{argmin}} \quad \sum_{i=1}^n \min_{j=1} \left\| \mathbf{y}^{(i)} - \mathbf{v}_j \right\|_2^2 .$$

To illustrate the relationship between this formulation and our setup (with $C = \Delta^q$), suppose we specify the linear map $A \in \mathbb{R}^{d\times q}$ in (3.2) in terms of its columns as $A = [\mathbf{a}_1|\ldots|\mathbf{a}_q]$. Then the optimization problem (3.2) can be reformulated as

$$\underset{\{\mathbf{a}_j\}_{j=1}^q \subset \mathbb{R}^d}{\operatorname{argmin}} \quad \sum_{i=1}^n \left(y^{(i)} - \max_j \left\langle \mathbf{a}_j, \mathbf{u}^{(i)} \right\rangle \right)^2 . \tag{3.3}$$

One can view (3.3) as a problem of computing a partition of the collection of support function measurements into q disjoint clusters, with each cluster being assigned to an extreme point that minimizes the squared-loss error. In Section 3.4 we highlight the analogies between our algorithms for computing $\hat{\mathcal{K}}$ and the widely used Lloyd's algorithm for K-means clustering [93].

3.1.4 Outline

In Section 3.2 we discuss the geometric and algebraic aspects of the optimization problem (3.2), which serve as the foundation for the subsequent statistical analysis in Section 3.3. Throughout both of these sections, we describe several examples that provide additional insight into the mathematical development. We describe algorithms for (3.2) in Section 3.4, and we demonstrate the application of these methods in a range of numerical experiments in Section 3.5. We conclude with a discussion of future directions in Section 3.6.

3.2 Projections of Structured Convex Sets

In this section, we provide the necessary background to analyze our approach. Throughout this paper, we assume that the underlying set $\mathcal{K}^\star \subset \mathbb{R}^d$ is compact and convex with non-empty interior. We also assume that the lifting set $C \subset \mathbb{R}^q$, $q \geq d$, is compact and convex. In addition, we require C to satisfy the property that $0 \notin \operatorname{aff}(C)$, and that the cone generated by C is full dimensional – the first stipulation allows us to express translations of projections of C in our framework, and the second stipulation ensures that the set of linear transformations preserving C are invertible (we elaborate on this point later).

Notation. Given a convex set $C \subset \mathbb{R}^q$, we denote $\|A\|_{C,2} := \sup_{\mathbf{x}\in C} \|A\mathbf{x}\|_2$. We denote $\mathcal{B}_{\|\cdot\|}(\mathbf{x}) := \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| \leq 1\}$ to be the unit $\|\cdot\|$-ball centered at $\mathbf{x}$, and we denote $\|\cdot\|_F$ to be the Euclidean-norm over the space of matrices. Given a point $\mathbf{a} \in \mathbb{R}^q$ and a subset $\mathfrak{B} \subseteq \mathbb{R}^q$, we denote $\operatorname{dist}_{\|\cdot\|}(\mathbf{a}, \mathfrak{B}) := \inf_{\mathbf{b}\in\mathfrak{B}} \|\mathbf{a} - \mathbf{b}\|$. Last, given any two subsets $\mathfrak{A}, \mathfrak{B} \subset \mathbb{R}^q$, we denote the Hausdorff distance $d_H(\mathfrak{A}, \mathfrak{B}) := \inf_{t\geq 0}\{t : \mathfrak{A} \subseteq \mathfrak{B} + t\mathcal{B}_{\|\cdot\|_{\ell_2}}(0), \mathfrak{B} \subseteq \mathfrak{A} + t\mathcal{B}_{\|\cdot\|_{\ell_2}}(0)\}$.

Probabilistic Model for Support Function Measurements. We assume that the support function evaluations are *measurement pairs* – these are pairs of the form $(\mathbf{u}, y) \in \mathcal{S}^{d-1} \times \mathbb{R}$, where $\mathbf{u}$ denotes the direction at which the support function is evaluated, and y denotes the noisy support function evaluation – are independently and identically distributed according to the model:

$$P_{\mathcal{K}^\star}: \quad y = h_{\mathcal{K}^\star}(\mathbf{u}) + \varepsilon.$$

Here, $\mathbf{u} \in \mathcal{S}^{d-1}$ is a random vector distributed uniformly at random (u.a.r.) over the unit sphere, ε is a centered random variable with variance σ^2 (i.e. $\mathbb{E}[\varepsilon] = 0$, and $\mathbb{E}[\varepsilon^2] = \sigma^2$), and $\mathbf{u}$ and ε are independently distributed.

In the following, we describe the geometrical, algebraic, and analytical aspects of our procedure. The proofs of all results we state are straightforward. To simplify the exposition in this section, we defer these proofs to the Appendix.

3.2.1 Geometrical Aspects of Our Method

We quantify the difference between pairs of convex sets in terms of the L_p metric applied to their respective support functions. Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be a pair of compact convex sets in $\mathbb{R}^d$, and let $h_{\mathcal{K}_1}(\cdot)$ and $h_{\mathcal{K}_1}(\cdot)$ be the respective support functions of these sets. We denote

$$\rho_p(\mathcal{K}_1, \mathcal{K}_2) = \left(\int_{\mathcal{S}^{d-1}} |h_{\mathcal{K}_1}(\mathbf{u}) - h_{\mathcal{K}_2}(\mathbf{u})|^p \, \mu(d\mathbf{u}) \right)^{1/p}, \quad 1 \leq p < \infty, \tag{3.4}$$

where the integral is performed with respect to the Lebesgue measure over $\mathcal{S}^{d-1}$; as usual, we denote $\rho_\infty(\mathcal{K}_1, \mathcal{K}_2) = \max_{\mathbf{u}} |h_{\mathcal{K}_1}(\mathbf{u}) - h_{\mathcal{K}_2}(\mathbf{u})|$. In Section 3.3.1, we prove our convergence guarantees in terms of the L_p-metric.

The ρ_p-metric represents an important class of distance measures over convex sets. For instance, it features prominently in the literature of approximating convex sets as polytopes [25]. In addition, the specific instance of $p = \infty$ coincides with the Hausdorff distance [124].

A basic question we seek to address in Section 3.3.1 is to characterize the settings in which the sequences of estimators obtained using our method converge, and the limit to which these sequences converge to. As the estimators we consider are computed by minimizing an empirical loss function, a natural strategy is to understand the minimizers of the loss function at the *population level*. Let Q be a probability measure over the pairs $(\mathbf{u}, y) \in S^{d-1} \times \mathbb{R}$. We denote the loss function with respect

to Q as follows:

$$\Phi_C(A, Q) := \mathbb{E}_Q\{(h_C(A'\mathbf{u}) - y)^2\}.$$

First, we note that $\Phi_C(\cdot, Q)$ is continuous if y is Q-integrable:

Proposition 3.2.1 *Let Q be any probability distribution over the measurement pairs $(\mathbf{u}, y)$ satisfying $\mathbb{E}_Q[|y|] < \infty$. Then the function $A \mapsto \Phi_C(A, Q)$ is continuous at every A.*

Next, we denote the set of minimizers of the loss function at the population level $\Phi_C(\cdot, P_{\mathcal{K}^\star})$ by $\mathcal{M}_{\mathcal{K}^\star,C}$:

$$\mathcal{M}_{\mathcal{K}^\star,C} := \operatorname*{argmin}_A \Phi_C(A, P_{\mathcal{K}^\star}).$$

The following result states a series of properties about the set $\mathcal{M}_{\mathcal{K}^\star,C}$. Crucially, it shows that $\mathcal{M}_{\mathcal{K}^\star,C}$ characterizes optimal approximations of $\mathcal{K}^\star$ as projections of C:

Proposition 3.2.2 *The set $\mathcal{M}_{\mathcal{K}^\star,C}$ is compact and non-empty. Moreover, we have*

$$\hat{A} \in \mathcal{M}_{\mathcal{K}^\star,C} \quad \Leftrightarrow \quad \hat{A} \in \operatorname*{argmin}_{A \in L(\mathbb{R}^q, \mathbb{R}^d)} \rho_2(A(C), \mathcal{K}^\star).$$

It follows from Proposition 3.2.2 that an optimal approximation of $\mathcal{K}^\star$ as the projection of C always exists. In Section 3.3.1, we show that the estimators obtained using our method converge to such a set if it is also unique.

Example. *Suppose $\mathcal{K}^\star$ is the regular q-gon in $\mathbb{R}^2$, and C is the free spectrahedron O^2. Then $\mathcal{M}_{\mathcal{K}^\star,C}$ uniquely specifies ℓ_2-ball.*

Example. *Suppose $\mathcal{K}^\star$ is the unit ℓ_2-ball in $\mathbb{R}^2$, and C is the simplex Δ^q. Then $\mathcal{M}_{\mathcal{K}^\star,C}$ specifies a centered regular q-gon, but with an unspecified rotation.*

In light of our previous remark, a natural question to consider is to obtain a full characterization of the settings in which $\mathcal{M}_{\mathcal{K}^\star,C}$ defines a unique set. Unfortunately, such an undertaking is difficult as the set $\mathcal{M}_{\mathcal{K}^\star,C}$ is highly dependent on the sets $\mathcal{K}^\star$ and C. Based on Proposition 3.2.2, we can provide a simple sufficient condition under which $\mathcal{M}_{\mathcal{K}^\star,C}$ defines a unique set:

Corollary 3.2.3 *Suppose we have $\mathcal{K}^\star = A^\star(C)$ for some $A^\star \in L(\mathbb{R}^q, \mathbb{R}^d)$. Then the set of minimizers $\mathcal{M}_{\mathcal{K}^\star,C}$ uniquely define $\mathcal{K}^\star$; i.e., $\mathcal{K}^\star = A(C)$ for all $A \in \mathcal{M}_{\mathcal{K}^\star,C}$.*

In our earlier example where $\mathcal{K}^\star$ is the unit ℓ_2-ball in $\mathbb{R}^2$ and C is the simplex Δ^q, the set $\mathcal{M}_{\mathcal{K}^\star,C}$ contains multiple orbits because of non-trivial symmetries in $\mathcal{K}^\star$ – many sets that one encounters in practice do not contain such symmetries, in which case we expect $\mathcal{M}_{\mathcal{K}^\star,C}$ to define a unique set.

3.2.2 Algebraic Aspects of Our Method

Next, we introduce the necessary tools to view our geometric problem of reconstructing a set as an algebraic task of recovering a linear map.

We begin by describing the identifiability issues that arise with such an approach. Given a compact convex C, let g be the linear transformation that preserves C; i.e., $g(C) = C$. Then the linear map defined by Ag specifies the same convex set:

$$[Ag](C) = A(g(C)) = A(C) = \mathcal{K}.$$

As such, every projection map $A \in L(\mathbb{R}^q, \mathbb{R}^d)$ is a member of the equivalence relation defined by:

$$A \sim Ag, \quad g \in \mathrm{Aut}(C). \tag{3.5}$$

Here $\mathrm{Aut}(C)$ denotes the subset of linear transformations that preserve C. Since the cone generated by C is full dimensional, every element of $\mathrm{Aut}(C)$ must be an invertible matrix, and hence $\mathrm{Aut}(C)$ forms a subgroup of $\mathrm{GL}(q, \mathbb{R})$. In particular, the equivalence class $\mathfrak{O}_C(A) := \{Ag : g \in \mathrm{Aut}(C)\}$ specified by (3.5) is also the orbit of $A \in L(\mathbb{R}^q, \mathbb{R}^d)$ under group action by $\mathrm{Aut}(C)$. A further consequence of C being compact convex is that $\mathrm{Aut}(C)$ forms a compact matrix Lie group (see, e.g., Corollary 3.6 of [73]), and hence $\mathfrak{O}_C(A)$ is also a smooth manifold.

It follows that the space of projection maps $L(\mathbb{R}^q, \mathbb{R}^d)$ can be viewed as a union of orbits $\mathfrak{O}_C(A)$. An important property of the Hausdorff distance is that it defines a metric over collections of non-empty compact sets, and hence the collection of all orbits $\{\mathfrak{O}_C(A)\}_{A \in L(\mathbb{R}^q, \mathbb{R}^d)}$ endowed with the Hausdorff distance defines a metric space. In our subsequent analysis in Section 3.3, it is useful to view our set regression instance as one of recovering an orbit

Notice that the function $\Phi_C(\cdot, P_{\mathcal{K}^\star})$ is invariant over orbits of A: for every $g \in \mathrm{Aut}(C)$, we have $h_C(A'\mathbf{u}) = h_C((Ag)'\mathbf{u})$. It follows immediately that the set of minimizers $\mathcal{M}_{\mathcal{K}^\star,C}$ must also be a union of orbits. In Section 3.3.1, we show that the sequence of orbits corresponding to the minimizers of the empirical loss approach $\mathcal{M}_{\mathcal{K}^\star,C}$ as the number of measurements increase. Our results in Sections 3.3.2 and 3.3.3 consider the specialized setting where $\mathcal{M}_{\mathcal{K}^\star,C}$ is a single orbit, in which case the

sequence of orbits also converges to $\mathcal{M}_{\mathcal{K}^\star,C}$. It is straightforward to see that such a condition is satisfied if $\mathcal{K}^\star$ is a polytope with q extreme points, and we choose the lifting set C to be the simplex Δ^q in q-dimensions. The situation for projections of the free spectrahedron is more delicate, as the number of extreme points may be infinite. One simple instance in which $\mathcal{M}_{\mathcal{K}^\star,C}$ is a single orbit is when $\mathcal{K}^\star$ is the projection of O^q under a linear map A, and A maps O^q bijectively to $\mathcal{K}^\star$. The following result provides a procedure to construct further examples of convex sets that are representable as projections of the free spectrahedron, and where the set of minimizers $\mathcal{M}_{\mathcal{K}^\star,C}$ is a single orbit:

Proposition 3.2.4 *Let $\{\mathcal{K}_i\}_{i=1}^k \subset \mathbb{R}^d$ be a collection of sets with the property that $\mathcal{K}_i$ is expressible as the projection of O^{q_i}, and that $\mathcal{M}_{\mathcal{K}_i,O^{q_i}}$ is a single orbit. Let $\mathcal{K} = \mathrm{conv}(\cup_i \{\mathcal{K}_i\}_{i=1}^k)$. Suppose that $\mathcal{K}_i$ are exposed faces of $\mathcal{K}$. Then $\mathcal{K}$ is expressible as the projection of O^q, where $q = \sum_i q_i$. In addition, $\mathcal{M}_{\mathcal{K},O^q}$ is a single orbit.*

3.2.3 Analytical Aspects of Our Method

Third, we state the main derivative computations in our paper. These are useful for describing our examples in Section 3.3.2, and our algorithms in Section 3.4.

Given a compact convex C, the support function $h_C(\cdot)$ is differentiable at $\mathbf{x}$ if and only if $\mathrm{argmax}_{\mathbf{g}\in\mathbb{R}^q} \langle \mathbf{g}, \mathbf{x}\rangle$ is a singleton, and in which case, the derivative is given by $\mathrm{argmax}_{\mathbf{g}\in\mathbb{R}^q} \langle \mathbf{g}, \mathbf{x}\rangle$ (see Corollary 1.7.3 in [124]). We denote the derivative of h_C at $\mathbf{x}$ by $\boldsymbol{e}_C(\mathbf{x}) := \nabla_{\mathbf{x}}(h_C(\mathbf{x}))$.

Example. *Suppose $C = \Delta^q \subset \mathbb{R}^q$ is the simplex. The function $h_C(\cdot)$ is the maximum entry of the input vector, and is differentiable if and only if the maximum is unique, in which case the derivative $\boldsymbol{e}_C(\cdot)$ is the corresponding standard basis vector.*

Example. *Suppose $C = O^p \subset \mathbb{S}^p$ is the free spectrahedron. The function $h_C(\cdot)$ is the largest eigenvalue of the input matrix, and is differentiable if and only if the largest eigenvalue has multiplicity one, in which case the derivative $\boldsymbol{e}_C(\cdot)$ is the corresponding rank-one unit-norm positive definite matrix.*

The following result computes the first derivative of $\Phi_C(\cdot, P_{\mathcal{K}^\star})$.

Proposition 3.2.5 *Let Q be a probability distribution over the measurement pairs $(\mathbf{u}, y)$, and suppose that $\mathbb{E}_Q\{y^2\} < \infty$. Suppose that $A \in L(\mathbb{R}^q, \mathbb{R}^d)$ is a linear*

map such that $h_C(\cdot)$ is differentiable at $A'\mathbf{u}$ for Q-a.e. $\mathbf{u}$. Then $A \mapsto \Phi_C(A, Q)$ is differentiable at A with derivative $2\mathbb{E}_Q\{(h_C(A'\mathbf{u}) - y)\mathbf{u} \otimes \boldsymbol{e}_C(A'\mathbf{u})\}$.

3.3 Main Results

In this section, we describe the main performance guarantees associated with using our approach. We denote $P_{\mathcal{K}^\star} := \frac{1}{n}\sum_{i=1}^n \delta_{(\mathbf{u},y)=(\mathbf{u}^{(i)},y^{(i)})}$, where $(\mathbf{u}^{(i)}, y^{(i)}) \sim P_{\mathcal{K}^\star}$, to be the empirical measure corresponding to drawing n i.i.d. measurements from the model $P_{\mathcal{K}^\star}$, and we denote the estimator $\hat{\mathcal{K}}_n(C) := \hat{A}_n(C)$, where $\hat{A}_n \in \operatorname{argmin}_A \Phi_C(A, P_{n,\mathcal{K}^\star})$.

3.3.1 Strong Consistency

Our main result describing strong consistency shows that the sequence of estimators $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^\infty$ obtained using our method converges to the optimal ρ_2 approximation of $\mathcal{K}^\star$ as a projection of C, provided that such an approximation is unique:

Theorem 3.3.1 *Given a compact convex $\mathcal{K}^\star \subset \mathbb{R}^d$, and the choice of any compact convex $C \subset \mathbb{R}^q$ as the lifting set, let $\{\hat{A}_n\}_{n=1}^\infty$, $\hat{A}_n \in \operatorname{argmin}_A \Phi_C(A, P_{n,\mathcal{K}^\star})$, be a sequence of minimizers of the empirical loss function, and let $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^\infty$, $\hat{\mathcal{K}}_n(C) = \hat{A}_n(C)$, be the corresponding sequence of estimators of $\mathcal{K}^\star$. We have $\operatorname{dist}(\hat{A}_n, \mathcal{M}_{\mathcal{K}^\star,C}) \to 0$ a.s., and $\inf_{A\in\mathcal{M}_{\mathcal{K}^\star,C}} d_H(\mathfrak{D}_C(\hat{A}_n), \mathfrak{D}_C(A)) \to 0$ a.s.. In particular, if $\mathcal{M}_{\mathcal{K}^\star,C}$ specifies a unique set, i.e. there exists $\hat{\mathcal{K}} \subset \mathbb{R}^d$ such that $\hat{\mathcal{K}} = A(C)$ for all $A \in \mathcal{M}_{\mathcal{K}^\star,C}$, then $\rho_p(\hat{\mathcal{K}}_n(C), \hat{\mathcal{K}}) \to 0$ a.s.. Furthermore, if $\mathcal{K}^\star = A^\star(C)$ for some linear map $A^\star \in L(\mathbb{R}^q, \mathbb{R}^d)$, then $\rho_p(\hat{\mathcal{K}}_n(C), \mathcal{K}^\star) \to 0$ a.s..*

The above result is valid for all ρ_p metrics, where $1 \le p \le \infty$.

In settings where $\mathcal{M}_{\mathcal{K}^\star,C}$ defines multiple sets, our result implies a weaker notion of convergence: Given any $\epsilon > 0$, there exists a n_0 such that $\rho_p(\hat{\mathcal{K}}_n(C), A_n(C)) < \epsilon$ for some $A_n \in \mathcal{M}_{\mathcal{K}^\star,C}$, and all $n \ge n_0$, a.s.. In other words, although the sequence $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^\infty$ may not converge, we are guaranteed that its elements eventually approximate some member of the sets specified by $\mathcal{M}_{\mathcal{K}^\star,C}$ (not necessarily the same set for every element) to arbitrary precision.

Example. *Suppose $\mathcal{K}^\star$ is the unit ℓ_2-ball in $\mathbb{R}^2$, and $C = \Delta^q$. The optimal ρ_2 approximation is the regular q-gon with an unspecified rotation. The sequence $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^\infty$ does not have a limit; rather, there is a sequence $\{g_n\}_{n=1}^\infty \subset SO(2)$ such that $g_n\hat{\mathcal{K}}_n(C)$ converges to $\mathcal{K}^\star$ a.s..*

The proof of Theorem 3.3.1 comprises two parts. First, we show that there exists a sufficiently large ball $\mathcal{U} \subset L(\mathbb{R}^q, \mathbb{R}^d)$ such that $\hat{A}_n \in \mathcal{U}$ as $n \to \infty$ eventually a.s.. Second, we appeal to a uniform strong law of large numbers (SLLN), which we state in the following, and prove in the Appendix. The structure of our proof is similar to that of a corresponding result for K-means clustering (see the main theorem in [112]).

Lemma 3.3.2 (Uniform SLLN) *Let $\mathcal{U} \subset L(\mathbb{R}^q, \mathbb{R}^d)$ be bounded. Let Q be a probability distribution over the measurement pairs $(\mathbf{u}, y)$ satisfying $\mathbb{E}_Q\{y^2\} < \infty$, and let Q_n be the empirical measure corresponding to drawing n i.i.d. samples from the distribution Q. Consider the collection of functions $\mathfrak{G} := \{(h_C(A'\mathbf{u}) - y)^2\}_{A \in \mathcal{U}}$ in the variables $(\mathbf{u}, y)$. Then*

$$\sup_{g \in \mathfrak{G}} \left| \mathbb{E}_{Q_n}\{g\} - \mathbb{E}_Q\{g\} \right| \to 0 \quad as \quad n \to \infty \quad a.s..$$

The proof of Lemma 3.3.2 is located in the Appendix.

Proof of Theorem 3.3.1. First we show that $\text{dist}(\hat{A}_n, \mathcal{M}_{\mathcal{K}^\star, C}) \to 0$ a.s.. Pick any $c \in (0,1)$, and define $\mathcal{G}_{\mathbf{v},r,c} := \{(\mathbf{u}, y) : \langle \mathbf{v}, \mathbf{u} \rangle \geq c, |y| \leq rc/2\}$, for all $\mathbf{v} \in \mathcal{S}^{d-1}$, $r \geq 0$. Define $s_c = \mathbb{P}\{(\mathbf{u}, y) : \langle \mathbf{v}, \mathbf{u} \rangle \geq c\}$, and note that $s_c > 0$ since $c \in (0,1)$. Note that $\mathbb{P}\{\mathcal{G}_{\mathbf{v},r,c}\}$ is a continuous function in $\mathbf{v}$. Since $\mathbb{E}\{|y|\} < \infty$, we have $\mathbb{P}\{\mathcal{G}_{\mathbf{v},r,c}\} \uparrow s_c$ as $r \to \infty$. Hence $\{\{\mathbb{P}\{\mathcal{G}_{\mathbf{v},r,c}\}\}_{\mathbf{v} \in \mathcal{S}^{d-1}}\}_{r \geq 0}$ is a sequence of collections of continuous functions in $\mathbf{v}$ that converge pointwise to the constant function s_c as $r \to \infty$. Since the domain of these functions is $\mathcal{S}^{d-1}$, which is compact, the convergence is also uniform. Thus we can pick r sufficiently large so that $(r^2c^2/4)\mathbb{P}\{\mathcal{G}_{\mathbf{v},r,c}\} > \Phi_C(0, P_{\mathcal{K}^\star})$ uniformly over all $\mathbf{v} \in \mathcal{S}^{d-1}$.

We claim that $\hat{A}_n \in r\mathcal{B}_{\|\cdot\|_{C,2}}(0)$ eventually as $n \to \infty$ a.s.. Suppose on the contrary that this does not hold. Then $\hat{A}_n \notin r\mathcal{B}_{\|\cdot\|_{C,2}}(0)$ i.o.. For every $\hat{A}_n \notin r\mathcal{B}_{\|\cdot\|_{C,2}}(0)$ there is $\mathbf{x}_n \in C$ such that $\|\hat{A}_n \mathbf{x}_n\| > r$. Consider the sequence of unit-norm vectors $\hat{A}_n \mathbf{x}_n / \|\hat{A}_n \mathbf{x}_n\|_2$ over the set of indices n such that $\hat{A}_n \notin r\mathcal{B}_{\|\cdot\|_{C,2}}(0)$. Because the unit-sphere is compact, we can pick a convergent subsequence whose limit point is $\mathbf{v} \in \mathcal{S}^{d-1}$. Then

$$\begin{aligned}
\limsup_n \Phi_C(\hat{A}_n, P_{n,\mathcal{K}^\star}) &\geq \limsup_n \mathbb{E}_{P_{n,\mathcal{K}^\star}}\{\mathbf{1}(\mathcal{G}_{\mathbf{v},r,c})(h_C(\hat{A}_n'\mathbf{u}) - y)^2\} \\
&\geq \lim_n \mathbb{P}_{P_{n,\mathcal{K}^\star}}\{\mathcal{G}_{\mathbf{v},r,c}\} r^2c^2/4 = \mathbb{P}_P\{\mathcal{G}_{\mathbf{v},r,c}\} r^2c^2/4 > \Phi_C(0, P_{\mathcal{K}^\star}).
\end{aligned}$$

By the SLLN we have $\Phi_C(0, P_{n,\mathcal{K}^\star}) = \mathbb{E}_{P_{n,\mathcal{K}^\star}}\{y^2\} \to \mathbb{E}_P\{y^2\} = \Phi_C(0, P_{\mathcal{K}^\star})$ a.s.. Then $\Phi_C(\hat{A}_n, P_{n,\mathcal{K}^\star}) > \Phi_C(0, P_{n,\mathcal{K}^\star})$ i.o., which contradicts the optimality of $\hat{A}_n$.

We proceed to conclude that $\text{dist}(\hat{A}_n, \mathcal{M}_{\mathcal{K}^\star,C}) \to 0$ a.s.. Let $\hat{A} \in \mathcal{M}_{\mathcal{K}^\star,C}$ be arbitrary. First by the optimality of $\hat{A}_n$ we have $\Phi_C(\hat{A}_n, P_{n,\mathcal{K}^\star}) \leq \Phi_C(\hat{A}, P_{n,\mathcal{K}^\star})$ for all n. Second by Lemma 3.3.2 with the choice of $Q = P_{\mathcal{K}^\star}$, we have $\Phi_C(\hat{A}_n, P_{n,\mathcal{K}^\star}) \to \Phi_C(\hat{A}_n, P_{\mathcal{K}^\star})$, and $\Phi_C(\hat{A}, P_{n,\mathcal{K}^\star}) \to \Phi_C(\hat{A}, P_{\mathcal{K}^\star})$, both uniformly and a.s.. By combining the previous two outcomes we have $\Phi_C(\hat{A}_n, P_{\mathcal{K}^\star}) < \Phi_C(\hat{A}, P_{\mathcal{K}^\star}) + \eta$ eventually, for all $\eta > 0$. Third by Proposition 3.2.1 $\Phi_C(\cdot, P_{\mathcal{K}^\star})$ is continuous, and hence $\hat{A}_n \in \mathcal{U}$ eventually a.s, for any neighborhood $\mathcal{U}$ of $\mathcal{M}_{\mathcal{K}^\star,C}$. Thus $\text{dist}(\hat{A}_n, \mathcal{M}_{\mathcal{K}^\star,C}) \to 0$ a.s..

We conclude the remaining assertions. Note that since $\text{Aut}(C)$ is compact, we may bound $\sigma_{\max}(g)$ uniformly over all $g \in \text{Aut}(C)$ by some $c > 0$. Here, $\sigma_{\max}(\cdot)$ is the spectral norm. Given $\epsilon > 0$, let $\bar{A} \in \mathcal{M}_{\mathcal{K}^\star,C}$ be such that $\|\hat{A}_n - \bar{A}\|_F < \text{dist}(\mathfrak{O}_C(\hat{A}_n), \mathcal{M}_{\mathcal{K}^\star,C}) + \epsilon$. Then, for every $g \in \text{Aut}(C)$, we have $\|\hat{A}_n g - \bar{A} g\|_F \leq \sigma_{\max}(g)(\text{dist}(\mathfrak{O}_C(\hat{A}_n), \mathcal{M}_{\mathcal{K}^\star,C}) + \epsilon) \leq c(\text{dist}(\mathfrak{O}_C(\hat{A}_n), \mathcal{M}_{\mathcal{K}^\star,C}) + \epsilon)$. This implies that $d_H(\mathfrak{O}_C(\hat{A}_n), \mathfrak{O}_C(\bar{A})) \leq c(\text{dist}(\mathfrak{O}_C(\hat{A}_n), \mathcal{M}_{\mathcal{K}^\star,C}) + \epsilon)$. By noting that ϵ is arbitrary, it follows that $\inf_{A \in \mathcal{M}_{\mathcal{K}^\star,C}} d_H(\mathfrak{O}_C(\hat{A}_n), \mathfrak{O}_C(A)) \to 0$ a.s..

Now suppose that $\hat{A}(C)$ defines the same set for all $\hat{A} \in \mathcal{M}_{\mathcal{K}^\star,C}$. Then $h_C(\hat{A}'\mathbf{u})$ defines the same function for all $\hat{A} \in \mathcal{M}_{\mathcal{K}^\star,C}$. We have $\text{dist}(\hat{A}_n, \mathcal{M}_{\mathcal{K}^\star,C}) \to 0$ a.s., and thus $h_C(\hat{A}_n'\mathbf{u}) \to h_C(\hat{A}'\mathbf{u})$ pointwise. Note that $h_C(A'\mathbf{u})$ is a continuous function in $\mathbf{u}$ defined over a compact domain $\mathcal{S}^{d-1}$. Thus $h_C(\hat{A}_n'\mathbf{u}) \to h_C(\hat{A}'\mathbf{u})$ uniformly. Hence $\rho_p(\hat{\mathcal{K}}_n(C), \hat{\mathcal{K}}) \to 0$ a.s.. The last assertion follows from Corollary 3.2.3. □

3.3.2 Asymptotic Normality

Our second result characterizes the limiting distribution of the estimators $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^\infty$.

We begin our discussion by describing the nature of our result. Recall that each set $\hat{\mathcal{K}}_n(C) = \hat{A}_n(C)$ is specified by a projection map $\hat{A}_n \in L(\mathbb{R}^q, \mathbb{R}^d)$. The manner in which we describe the behavior of the set $\hat{\mathcal{K}}_n(C)$ is by characterizing the asymptotic distribution of $\hat{A}_n$. We do so by selecting a minimizer $\hat{A} \in \mathcal{M}_{\mathcal{K}^\star,C}$ to serve as a reference point. Following which, we define a sequence $\{\tilde{A}_n\}_{n=1}^\infty$, where each $\tilde{A}_n$ is a member from the equivalence class corresponding to $\hat{A}_n$ (see (3.5)) that is closest to $\hat{A}$ in the Euclidean distance. Our main result in this subsection shows that, under conditions which we describe next, the sequence $\{\sqrt{n}(\tilde{A}_n - \hat{A})\}_{n=1}^\infty$ is asymptotically normal. In our subsequent discussion, we also illustrate a series of examples in which we apply our result to describe the asymptotic behavior of the set $\hat{\mathcal{K}}_n(C)$.

Next, we describe the key ingredients required for our result. The first set of re-

quirements are non-degeneracy assumptions on the function $\Phi_C(\cdot, P_{\mathcal{K}^\star})$. First, we require the minimizers of $\Phi_C(\cdot, P_{\mathcal{K}^\star})$ to define a unique orbit. Following from the conclusions of Theorem 3.3.1, such an assumption guarantees that the sequence of orbits $\{\mathfrak{O}_C(\tilde{A}_n)\}_{n=1}^{\infty}$ converges, and in particular, it also guarantees that the sequence of linear maps $\{\tilde{A}_n\}_{n=1}^{\infty}$ converges, which is necessary in order to provide a Central Limit Theorem (CLT) type of characterization for $\{\tilde{A}_n\}_{n=1}^{\infty}$. Second, we require the function $\Phi_C(\cdot, P_{\mathcal{K}^\star})$ to be twice differentiable at $\hat{A}$, and whose second derivative is positive definite (modulo invariances due to $\mathrm{Aut}(C)$) – this is a non-degeneracy assumption that allows us to obtain a quadratic approximation of the function $\Phi_C(\cdot, P_{\mathcal{K}^\star})$ around $\hat{A}$, and subsequently compute first order approximations of the empirical minimizers $\tilde{A}_n$. The inclusion of these conditions have a geometric interpretation: they guarantee that images of extreme points of C under the projection map $\hat{A}_n$ – and by extension, the extreme points of $\hat{\mathcal{K}}_n(C)$ – converge in a "well-behaved" manner. In our subsequent discussion, we consider a specific instance where the above conditions are violated. We describe the behavior of $\hat{\mathcal{K}}_n(C)$ to illustrate the manner in which an asymptotic normality characterization may fail if the above conditions are not satisfied.

Our second ingredient is a property from empirical statistics known as *stochastic equicontinuity* – it is a notion of uniform continuity that is specialized to sequences of stochastic processes, and is useful for establishing a functional central limit theorem for minimizers of an empirical process:

Definition 3.3.3 (Stochastic equicontinuity) *Let $\{Z_n(t) : t \in \mathcal{T}\}_{n=1}^{\infty}$ be a sequence of stochastic processes whose index set $\mathcal{T}$ is equipped with a pseudo-metric $d(\cdot,\cdot)$.*[2] *The sequence $\{Z_n\}_{n=1}^{\infty}$ is said to be* stochastically equicontinuous *at a point $t_0 \in \mathcal{T}$ if for every $\eta > 0$ and $\epsilon > 0$ there exists a neighborhood $\mathcal{U}$ of t_0 such that*

$$\limsup_n \mathbb{P}_{\mathcal{K}^\star}\left\{\sup_{t\in\mathcal{U}} |Z_n(t) - Z_n(t_0)| > \eta\right\} < \epsilon.$$

We describe the specific sequence of stochastic process for which we need to demonstrate stochastic equicontinuity. First, let $E_{n,\mathcal{K}^\star}$ denote the signed measure:

$$E_{n,\mathcal{K}^\star} := \sqrt{n}(P_{n,\mathcal{K}^\star} - P_{\mathcal{K}^\star}).$$

[2] A pseudo-metric $d(\cdot,\cdot)$ is the same as a metric except that we do not require $d(x, y) = 0$ to imply that $x = y$.

Next, we define the following remainder term

$$\begin{aligned}&\lambda_C((\mathbf{u}, y), A, D)\\ :=\ &\frac{1}{\|D\|_{C,2}}\left((h_C((A+D)'\mathbf{u}) - y)^2 - (h_C(A'\mathbf{u}) - y)^2 - \langle \nabla_A((h_C(A'\mathbf{u}) - y)^2), D\rangle\right).\end{aligned} \tag{3.6}$$

Third, we define the stochastic process $\{\mathbb{E}_{E_n}\{\lambda_C(\cdot, \hat{A}, D)\} : D \in L(\mathbb{R}^q, \mathbb{R}^d)\}$, where $D \in L(\mathbb{R}^q, \mathbb{R}^d)$ denotes the index, and where the index space $L(\mathbb{R}^q, \mathbb{R}^d)$ is equipped with the seminorm $\|\cdot\|_{\mathcal{L}^2(P_{\mathcal{K}^\star})} := \mathbb{E}_{P_{\mathcal{K}^\star}}\{(\cdot)^2\}^{1/2}$. Here, note that $\mathbb{E}_{E_n}\{\lambda_C(\cdot, \hat{A}, D)\}$ is *not* deterministic because E_n is a random measure. To establish a CLT type of result, we need to show that the following sequence of stochastic processes is stochastically equicontinuous at $D = 0$:

$$\{\{\mathbb{E}_{E_n}\{\lambda_C(\cdot, \hat{A}, D)\} : D \in L(\mathbb{R}^q, \mathbb{R}^d)\}\}_{n=1}^{\infty}. \tag{3.7}$$

Observe that the remainder function λ_C is defined with respect to a choice of C, and hence stochastic equicontinuity of (3.7) is a property that depends on C. For many natural choices of lifting sets C arising in optimization, such a property is indeed satisfied – the following result shows that (3.7) is stochastically equicontinuous for the choice of C being the free spectrahedron:

Proposition 3.3.4 *Let $\lambda_C(\cdot, \hat{A}, D)$ be the remainder functions defined according to* (3.6) *with respect to choice of C being the free spectrahedron O^q. Then the sequence of stochastic processes $\{\{\mathbb{E}_{E_n}\{\lambda_C(\cdot, \hat{A}, D)\} : D \in L(\mathbb{R}^q, \mathbb{R}^d)\}\}_{n=1}^{\infty}$ where D is the index and $L(\mathbb{R}^q, \mathbb{R}^d)$ is equipped with the seminorm $\|\cdot\|_{\mathcal{L}^2(P_{\mathcal{K}^\star})}$, is stochastically equicontinuous at $D = 0$.*

The proof of Proposition 3.3.4 relies on a result by Stengle and Yukich showing that certain collection of sets admitting semialgebraic representations are Vapnik-Chervonenkis (VC) [136] – such a property is useful for establishing stochastic equicontinuity of sequences of stochastic processes. We prove the result, and outline extensions of Proposition 3.3.4 to other choices of sets C that admit semialgebraic descriptions in Section 3.3.2.1.

Our third ingredient is a mild structural assumption that requires the automorphisms of C to be isometries. We remark that such an assumption is not particularly restrictive, as many natural choices of lifting sets C satisfy such a requirement. For instance, the automorphism group of the simplex Δ^q is the set of all $q \times q$ permutation

matrices, and the automorphism group of the free spectrahedron $\mathcal{O}^p$ is the set of all operators $T : \mathbb{S}^p \to \mathbb{S}^p$, where $T(X) = WXW'$, and $W \in O(p)$ is orthogonal. The utility of such an assumption is that it allows us to conclude that $\tilde{A}_n - \hat{A}$ all lie in the normal space with respect to $\mathcal{M}_{\mathcal{K}^\star,C}$ at $\hat{A}$.

Theorem 3.3.5 *Let $\mathcal{K}^\star \subseteq \mathbb{R}^d$ be a compact convex set. Suppose that $C \subseteq \mathbb{R}^q$ is a compact convex set satisfying $\mathrm{Aut}(C) \lhd O(q)$, where $O(q)$ is the subgroup of $q \times q$ orthogonal matrices, and $\mathcal{M}_{\mathcal{K}^\star,C} = \mathfrak{O}_C(\hat{A})$, for some $\hat{A} \in L(\mathbb{R}^q, \mathbb{R}^d)$. Suppose that the map $A \mapsto \Phi_C(A, P)$ is twice differentiable at $\hat{A}$, and whose second derivative at $\hat{A}$, which we denote by Γ, is positive definite restricted to $\mathcal{N} := \mathcal{N}_{\mathcal{M}_{\mathcal{K}^\star,C}}(\hat{A})$; i.e., $\Gamma|_{\mathcal{N}} \succ 0$. In addition, suppose that the sequence of stochastic processes $\{\{\mathbb{E}_{E_n}\{\lambda_C(\cdot, \hat{A}, D)\} : D \in L(\mathbb{R}^q, \mathbb{R}^d)\}\}_{n=1}^\infty$ is stochastically equicontinuous at $D = 0$. Let $\{\hat{A}_n\}_{n=1}^\infty$, $\hat{A}_n \in \mathrm{argmin}_A \Phi_C(A, P_{n,\mathcal{K}^\star})$ be the sequence of minimizers of the empirical loss function, and denote $\tilde{A}_n \in \mathrm{argmin}_{A \in \mathfrak{O}_C(\hat{A}_n)} \|\hat{A} - A\|_F$. We have, restricted to $\mathcal{N}$,*

$$\sqrt{n}(\tilde{A}_n - \hat{A}) \xrightarrow{D} \mathcal{N}(0, (\Gamma|_{\mathcal{N}})^{-1}(\mathbb{E}_{P_{\mathcal{K}^\star}}\{\nabla \otimes \nabla|_{A=\hat{A}}\}|_{\mathcal{N}})(\Gamma|_{\mathcal{N}})^{-1}),$$

where $\nabla = \nabla_A(h_C(A'\mathbf{u} - y)^2)$.

As we noted earlier, an asymptotic characterization of $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^\infty$ may fail if the non-degeneracy conditions concerning $\Phi_C(\cdot, P_{\mathcal{K}^\star})$ are not satisfied – the following example illustrates our point:

Example. *Let $\mathcal{K}^\star := \{0\} \subset \mathbb{R}$ be a singleton, and let the noise $\{\varepsilon^{(i)}\}_{i=1}^n$ be i.i.d. centered Gaussian random variables with variance σ^2. Since $\mathcal{S}^1 \cong \{-1, 1\}$, the random variables $\mathbf{u}^{(i)}$ are ± 1 u.a.r.. Furthermore, since $h_{\mathcal{K}^\star}(\mathbf{u}) = 0$ for all $\mathbf{u}$, the support function measurements are simply $y^{(i)} = \varepsilon^{(i)}$, for all $1 \leq i \leq n$. Also, for any choice of C, the set $\mathcal{M}_{\mathcal{K}^\star,C} = \{0\} \subset L(\mathbb{R}^q, \mathbb{R})$ is a singleton consisting the zero map.*

First, we consider fitting $\mathcal{K}^\star$ with the choice of $C = \Delta^1 \subset \mathbb{R}^1$. Then $\hat{A}_n = \frac{1}{n}\sum_{i=1}^n \varepsilon^{(i)}\mathbf{u}^{(i)}$, from which it follows that $\sqrt{n}(\hat{A}_n - 0)$ is normally distributed with mean zero and variance σ^2 – this is in agreement with Theorem 3.3.5.

Second, we consider fitting $\mathcal{K}^\star$ with the choice of $C = \Delta^2 \subset \mathbb{R}^2$. Define the sets $\mathfrak{U}_- = \{i : \mathbf{u}_i = -1\}$, and $\mathfrak{U}_+ = \{i : \mathbf{u}_i = 1\}$, and define

$$\alpha_- = -\frac{1}{|\mathfrak{U}_-|}\sum_{i \in \mathfrak{U}_-} \varepsilon^{(i)} \quad \text{and} \quad \alpha_+ = \frac{1}{|\mathfrak{U}_+|}\sum_{j \in \mathfrak{U}_+} \varepsilon^{(j)}.$$

Then $\hat{\mathcal{K}}_n(C) = \{x : \alpha_- \leq x \leq \alpha_+\}$ *if* $\alpha_- \leq \alpha_+$, *and* $\hat{\mathcal{K}}_n(C) = \{\frac{1}{n}\sum_{i=1}^n \varepsilon^{(i)}\mathbf{u}^{(i)}\}$ *otherwise. Notice that* α_- *and* α_+ *have the same distribution, and hence* $\hat{\mathcal{K}}_n(C)$ *is a closed interval with non-empty interior w.p.* $1/2$, *and is a singleton w.p.* $1/2$. *For this particular instance, one can see that the linear map* $\hat{A}_n$ *does not satisfy an asymptotic normality characterization. A further computation reveals that the function* $\Phi_C(\cdot, P_{\mathcal{K}^\star})$ *is twice differentiable everywhere excluding the line* $\{(c,c) : c \in \mathbb{R}\}$*; in particular, it is not differentiable at the minimizer* $(0,0)$.

The above example is an instance where the function $\Phi_C(\cdot, P_{\mathcal{K}^\star})$ is not twice differentiable at $\hat{A}$. We remark that the manner in which an asymptotic characterization of $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^\infty$ fails in instances where $\mathcal{M}_{\mathcal{K}^\star,C}$ contains multiple orbits is also qualitatively similar.

The proof of Theorem 3.3.5 requires the following structural result, which follows as a consequence of $\mathrm{Aut}(C) \lhd O(q)$:

Lemma 3.3.6 *Let* $C \subset \mathbb{R}^q$ *be compact convex. Suppose that* $\mathrm{Aut}(C) \lhd O(q)$. *Let* $A, B \in L(\mathbb{R}^q, \mathbb{R}^d)$. *Define* $A_B \in \mathrm{argmin}_{X \in \mathfrak{O}_C(A)} \|B - X\|_F$, *and let* $g \in \mathrm{Aut}(C)$ *be such that* $A = A_B g$. *Then* $\|Bg - A\|_F = \|B - Ag^{-1}\|_F = \mathrm{dist}_{\|\cdot\|_F}(B, \mathfrak{O}_C(A))$, *and* $Bg - A \in \mathcal{N}_{\mathfrak{O}_C(A)}(A)$.

Proof of Lemma 3.3.6. The first series of equalities follow from g being an isometry. From the first optimality conditions of A_B, we have $B - A_B \in \mathcal{N}_{\mathfrak{O}_C(A)}(A_B)$. By applying the fact that g is an isometry, and that $\mathfrak{O}_C(A)$ is the orbit of A under group action of a subgroup of $O(d)$, we have $Bg - A \in \mathcal{N}_{\mathfrak{O}_C(A)}(A)$. □

Proof of Theorem 3.3.5. By combining Lemma 3.3.6 with Theorem 3.3.1, we have $\tilde{A}_n \to \hat{A}$ a.s., and $\tilde{A}_n - \hat{A} \in \mathcal{N}$, for all $n \geq 1$. Let $\mathcal{N}_{\mathrm{aff}}(\hat{A})$ be the affine subspace parallel to $\mathcal{N}$ containing $\hat{A}$. It follows that $\tilde{A}_n \in \mathcal{N}_{\mathrm{aff}}(\hat{A})$. The remainder of the proof is a direct application of Theorem 5 in Section VII of [113] to $\{\tilde{A}_n\}_{n=1}^\infty$, where $\{\tilde{A}_n\}_{n=1}^\infty$ is to be viewed as a sequence in $\mathcal{N}_{\mathrm{aff}}(\hat{A})$. We proceed to verify that the conditions of Theorem 5 hold. For completeness, we state the conditions of the result specialized to our setting. First, we require $\hat{A}$ to be an interior point of $\mathcal{N}_{\mathrm{aff}}(\hat{A})$ – this is trivially satisfied. Second, let $\tilde{\Phi}_C(\cdot, P_{\mathcal{K}^\star})$ be the function $\Phi_C(\cdot, P_{\mathcal{K}^\star})$ restricted onto $\mathcal{N}_{\mathrm{aff}}(\hat{A})$. We require the function $\tilde{\Phi}_C(\cdot, P_{\mathcal{K}^\star})$ to be twice differentiable at $\hat{A}$ with a non-singular derivative – this is satisfied by the assumption that $\Gamma|_{\mathcal{N}}$ is positive definite. Third, we require $\Phi_C(\tilde{A}_n, P_{n,\mathcal{K}^\star}) = o_p(n^{-1}) + \inf_A \Phi_C(A, P_{n,\mathcal{K}^\star})$ – this is satisfied as $\tilde{A}_n$ minimizes $\Phi_C(\cdot, P_{n,\mathcal{K}^\star})$. Fourth, we require the components of

$\nabla_A((h_C(\hat{A}'\mathbf{u}) - y)^2)$ to be in $\mathcal{L}^2(P_{\mathcal{K}^\star})$ – this follows from $h_C(\hat{A}'\mathbf{u})$, $\mathbf{u}$, and $\boldsymbol{e}_C(\hat{A}'\mathbf{u})$ are being uniformly bounded over $\mathbf{u} \in \mathcal{S}^{d-1}$, as well as $\mathbb{E}_{P_{\mathcal{K}^\star}}\{\varepsilon^2\} < \infty$. Fifth, we require the sequence $\{\mathbb{E}_{E_n}\{\lambda_C(\cdot, \hat{A}, D)\}\}_{n=1}^{\infty}$ indexed over $D \in \mathcal{N}$ to be stochastically equicontinuous at $D = 0$ – this follows the original assumption that the sequence $\{\mathbb{E}_{E_n}\{\lambda_C(\cdot, \hat{A}, D)\}\}_{n=1}^{\infty}$ indexed over $D \in L(\mathbb{R}^q, \mathbb{R}^d)$ is stochastically equicontinuous at $D = 0$. □

Specialization of Theorem 3.3.5. In the following, we specialize Theorem 3.3.5 to the setting where the underlying set $\mathcal{K}^\star = A^\star(C)$ is expressible as the projection of C. The inclusion of such an assumption allows us to compute an explicit expression of the second derivative of $\Phi_C(\cdot, P_{\mathcal{K}^\star})$:

Proposition 3.3.7 *Suppose that the underlying set $\mathcal{K}^\star = A^\star(C)$ for some $A^\star \in L(\mathbb{R}^q, \mathbb{R}^d)$. In addition, suppose that the function $h_C(\cdot)$ is continuously differentiable at $A^{\star\prime}\mathbf{u}$ for $P_{\mathcal{K}^\star}$-a.e. $\mathbf{u}$. Then the map $A \mapsto \Phi_C(A, P_{\mathcal{K}^\star})$ is twice differentiable at $A^\star$, and whose second derivative is the operator $\Gamma \in L(L(\mathbb{R}^q, \mathbb{R}^d), L^*(\mathbb{R}^q, \mathbb{R}^d))$ defined by*

$$\Gamma(D) = 2\mathbb{E}\left\{\langle \mathbf{u} \otimes \boldsymbol{e}_C(A^{\star\prime}\mathbf{u}), D\rangle \mathbf{u} \otimes \boldsymbol{e}_C(A^{\star\prime}\mathbf{u})\right\}. \tag{3.8}$$

The proof of Proposition 3.3.7 is a simple computation and is located in the Appendix. We remark that it is considerably more difficult to compute a general expression of the second derivative of $\Phi_C(\cdot, P_{\mathcal{K}^\star})$, and hence our result in Proposition 3.3.7 applies to a more restrictive setting compared to Proposition 3.2.5.

Our specialization of Theorem 3.3.5 is as follows:

Corollary 3.3.8 *Suppose that the conditions of Theorem 3.3.5 and Proposition 3.3.7 hold. Then, using the notation of Theorem 3.3.5, we have $\mathbb{E}\{\nabla \otimes \nabla|_{A=A^\star}\}|_{\mathcal{N}} = 2\sigma^2\Gamma|_{\mathcal{N}}$, with Γ given by* (3.8)*. In particular, the conclusion of Theorem 3.3.5 simplifies to $\sqrt{n}(\tilde{A}_n - \hat{A}) \xrightarrow{D} \mathcal{N}(0, 2\sigma^2(\Gamma|_{\mathcal{N}})^{-1})$.*

Proof of Corollary 3.3.8. We have $\nabla_A((h_C(A'\mathbf{u}) - y)^2)|_{A=A^\star} = -\varepsilon\mathbf{u} \otimes \boldsymbol{e}_C(A^{\star\prime}\mathbf{u})$, from which we have $\mathbb{E}\{\nabla \otimes \nabla\} = 2\sigma^2\Gamma$, and hence the result. □

Examples. In the following, we consider a series of examples in which we apply our method to reconstruct different instances of $\mathcal{K}^\star$. We interpret the conclusions of Theorem 3.3.5 to describe the distributional behavior of $\hat{\mathcal{K}}_n(C)$. To simplify

our discussion, we specify the choices of lifting set C and the projection map $A^\star \in L(\mathbb{R}^q, \mathbb{R}^d)$. In addition, our choices of C and $A^\star$ also satisfy the conditions of Corollary 3.3.8 (for the sake of brevity, we omit verifying that the conditions of Corollary 3.3.8 hold), which we apply to compute the second derivative of the map $A \mapsto \Phi_C(A, P_{\mathcal{K}^\star})$ at $A^\star$ denoted by Γ.

Our first and second examples consider instances of polyhedral $\mathcal{K}^\star$. We choose $C = \Delta^q$, where q is the number of extreme points of $\mathcal{K}^\star$. Under such a choice, the set $\mathcal{M}_{\mathcal{K}^\star, C}$ comprises linear maps $A \in L(\mathbb{R}^q, \mathbb{R}^d)$ whose columns are the extreme points of $\mathcal{K}^\star$. One can show that Γ is a block diagonal operator comprising q blocks of dimensions $d \times d$. Using these pieces of information together with Theorem 3.3.1, we conclude the following description about $\hat{\mathcal{K}}_n(C)$: (i) it is a polytope with q extreme points, (ii) every vertex of $\hat{\mathcal{K}}_n(C)$ is close to a unique vertex of $\mathcal{K}^\star$, (iii) the deviations (after scaling by a factor of $\sqrt{n}$) between every vertex-vertex pair are asymptotically normal with inverse covariance specified by a $d \times d$ block of Γ, and are pairwise-independent.

Example. *Let $\mathcal{K}^\star$ be the regular q-gon in $\mathbb{R}^2$ whose vertices are given by $\{\boldsymbol{v}_k\}_{k=0}^{q-1}$, where $\boldsymbol{v}_k := (\cos(2k\pi/q), \sin(2k\pi/q))'$. Let $\hat{\boldsymbol{v}}_{n,k}$ be the vertex of $\hat{\mathcal{K}}_n(C)$ closest to $\boldsymbol{v}_k$. The displacement $\sqrt{n}(\hat{\boldsymbol{v}}_{n,k} - \boldsymbol{v}_k)$ is a random vector, and is asymptotically normal with covariance $2\sigma^2 M_{k,k}^{-1}$, where:*

$$M_{k,k} = \frac{1}{q} I + \frac{1}{2\pi} \sin(2\pi/q) \begin{pmatrix} \cos(4k\pi/q) & \sin(4k\pi/q) \\ \sin(4k\pi/q) & -\cos(4k\pi/q) \end{pmatrix}.$$

The eigenvalues of $M_{k,k}$ are $1/q + (1/2\pi)\sin(2\pi/q)$ and $1/q - (1/2\pi)\sin(2\pi/q)$, and the corresponding eigenvectors are $(\cos(2k\pi/q), \sin(2k\pi/q))'$ and $(\sin(2k\pi/q), -\cos(2k\pi/q))'$ respectively. Consequently, the random vector $\hat{\boldsymbol{v}}_{n,k} - \boldsymbol{v}_k$ has magnitude $\approx \sigma\sqrt{q/n}$ in the direction $\boldsymbol{v}_k$, and has magnitude $\approx \sigma\sqrt{3q^3/\pi^2 n}$ in the direction $\boldsymbol{v}_k^\perp$. Figure 3.2 shows $\mathcal{K}^\star$, and the confidence intervals (ellipses) of the vertices of $\hat{\mathcal{K}}_n(C)$ for $q = 5$.

Example. *Let $\mathcal{K}^\star$ be the ℓ_∞-ball in $\mathbb{R}^d$. Let $\boldsymbol{v} \in \{(\pm 1, \ldots, \pm 1)'\}$ be a vertex of $\mathcal{K}^\star$, and let $\hat{\boldsymbol{w}}_{n,\boldsymbol{v}}$ denote the vertex of $\hat{\mathcal{K}}_n(C)$ closest to $\boldsymbol{v}$. The deviation $\hat{\boldsymbol{w}}_{n,\boldsymbol{v}} - \boldsymbol{v}$ is asymptotically normal with covariance $2\sigma^2 M_{\boldsymbol{v},\boldsymbol{v}}^{-1}$, where:*

$$M_{\boldsymbol{v},\boldsymbol{v}} = \frac{1}{2^d d}((1 - 2/\pi)I + (2/\pi)\boldsymbol{v}\boldsymbol{v}').$$

Hence it follows that $\hat{\boldsymbol{w}}_{n,\boldsymbol{v}} - \boldsymbol{v}$ is a random vector whose magnitude is $\approx \sigma 2^{(d+1)/2}(2/\pi + (1 - 2/\pi)/d)^{-1/2} n^{-1/2}$ in the direction of $\boldsymbol{v}$, and is $\approx \sigma 2^{(d+1)/2}(1 - 2/\pi)^{-1/2}\sqrt{d/n}$ in the subspace orthogonal to $\boldsymbol{v}$.

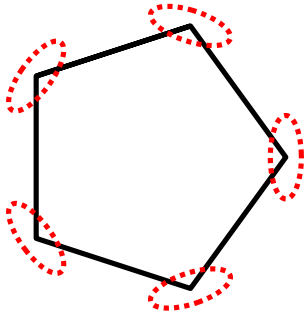


Figure 3.2: Estimating a regular 5-gon as the projection of Δ^5. In the large n limit, the estimator $\hat{\mathcal{K}}_n(C)$ is a 5-gon. The typical deviation of the vertices of $\hat{\mathcal{K}}_n(C)$ from that of the 5-gon (scaled by a factor of $\sqrt{n}$) is represented by the surrounding ellipses.

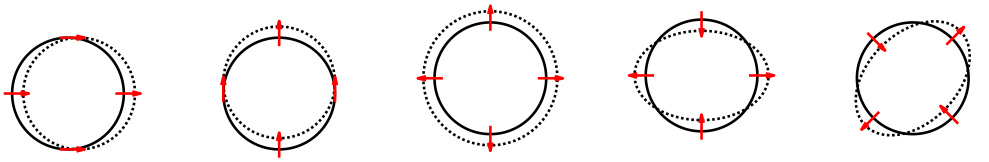
Our third and fourth examples consider instances of non-polyhedral $\mathcal{K}^\star$. Unlike the previous examples, our description of $\hat{\mathcal{K}}_n(C)$ requires a different interpretation of Corollary 3.3.8 that is suitable for sets with infinitely many extremal points. We express $\tilde{A}_n = A^\star + D_n$, where $\sqrt{n}D_n$ is a random linear map that is asymptotically normal with covariance $2\sigma^2(\Gamma|_{\mathcal{N}})^{-1}$, and where $\mathcal{N}$ denotes the normal space with respect to $\mathcal{M}_{\mathcal{K}^\star,C}$ at $A^\star$. We describe the behavior of $\hat{\mathcal{K}}_n(C)$ by characterizing the contribution of $D_n(C)$ as a perturbation to the set $\mathcal{K}^\star = A^\star(C)$.

Example. *Suppose $\mathcal{K}^\star = \mathcal{B}_{\|\cdot\|_2}(\mathbf{c})$ is the ℓ_2-ball in $\mathbb{R}^d$ with center $\mathbf{c}$. We choose $C := \{(1, \mathbf{v})' : \|\mathbf{v}\|_2 \leq 1\} \subset \mathbb{R}^{d+1}$, and $A^\star$ to be a linear map of the form $[\mathbf{c}\ Q] \in L(\mathbb{R}^{d+1}, \mathbb{R}^d)$, where $Q \in O(d)$ is any orthogonal matrix. Then $\Gamma|_{\mathcal{N}}$ is a self-adjoint operator with rank $q + \binom{q+1}{2}$. The eigenvectors of $\Gamma|_{\mathcal{N}}$ represent 'modes of oscillations', which we describe in greater detail.*

We begin with the case where $d = 2$ and $\mathbf{c} = 0$. Every perturbation $D_n(C)$ can be decomposed into 5 different modes (these exactly correspond to the eigenvectors of the operator $\Gamma|_{\mathcal{N}}$). To understand how these modes perturb $A^\star(C)$, we parameterize the extremal points of $A^\star(C)$ by $\{(\cos(\theta)\ \sin(\theta))'\}_{\theta\in[0,2\pi)}$. The contribution of each mode at the point $(\cos(\theta)\ \sin(\theta))'$ is a small perturbation in the direction $(1\,0)'$, $(0\,1)'$, $(\cos(\theta)\ \sin(\theta))'$, $(\cos(\theta)\ -\sin(\theta))'$, and $(\sin(\theta)\ \cos(\theta))'$ respectively – Figure 3.3 provides an illustration. Notice that the first and second modes represent oscillations of $\hat{\mathcal{K}}_n(C)$ about $\mathbf{c}$, the third mode represents dilation, and the fourth and fifth mode represent flattening.

The analysis for a general d is similar. Let $\{\mathbf{g} : \|\mathbf{g}\|_2 = 1\}$ denote the extreme points of $A^\star(C)$. First, there are q modes whose contributions are $\{(0, \ldots, g_i, \ldots, 0) : g_i = \mathbf{g}_i\}$, $1 \leq i \leq d$, and these represent oscillations about $\mathbf{c}$. Second, there are $\binom{q+1}{2}$

modes whose contributions are of the form $M\mathbf{g}$*, where* $\{M : M \in L(\mathbb{R}^d, \mathbb{R}^d), M = M'\}$*, and these represent higher dimensional analogs of flattening (of which dilation is a special case).*



(a) $(1\ \ 0)'$ (b) $(0\ \ 1)'$ (c) $(\cos(\theta)\ \ \sin(\theta))'$ (d) $(\cos(\theta)\ \ \sin(\theta))'$ (e) $(\sin(\theta)\ \ \cos(\theta))'$

Figure 3.3: Modes of oscillations for an estimate of the ℓ_2-ball in $\mathbb{R}^2$.

Example. *Let* $\mathcal{K}^\star$ *be the spectral norm ball in* $\mathbb{S}^2$*. The extreme points of* $\mathcal{K}^\star$ *comprise three connected components:* $\{I\}$*,* $\{-I\}$*, and* $\{UDU'\}_{U\in SO(2)}$*, where* D *is a diagonal matrix with entries* $(1,-1)$*. To simplify our discussion, we apply a scaled isometry to* $\mathcal{K}^\star$ *so that* $\{I\}$*,* $\{-I\}$*, and* $\{UDU'\}_{U\in SO(2)}$ *are mapped to* $\{(0,0,1)'\}$*,* $\{(0,0,-1)'\}$*, and* $\{(\cos(\theta),\sin(\theta),0)'\}_{\theta\in[0,2\pi)}$ *respectively. We proceed with our description with respect to the transformation. We choose*

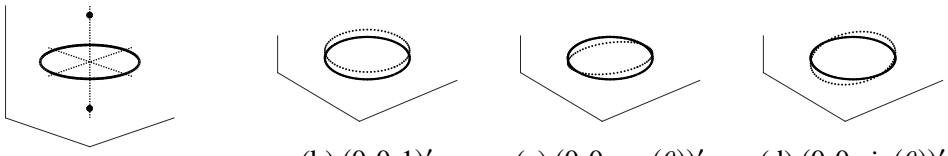
$$C := \{X : X \in O^4, X_{12} = X_{13} = X_{14} = X_{23} = X_{24} = X_{21} = X_{31} = X_{41} = X_{32} = X_{42} = 0\} \cong \mathbb{R}^2 \times \mathbb{S}^2, \tag{3.9}$$

and $A^\star$ *to be the map defined by* $A^\star(X) = (\langle A_1, X\rangle, \langle A_2, X\rangle, \langle A_3, X\rangle)'$*, where*

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

In the large n *limit,* $\hat{\mathcal{K}}_n(C)$ *is a convex set whose extremal points comprise a pair of points* $\mathcal{P}_1$ *and* $\mathcal{P}_2$ *near* $(0,0,1)'$ *and* $(0,0,-1)'$ *respectively, and an ellipse* $\mathcal{P}_3$ *near* $\{(\cos(\theta),\sin(\theta),0)'\}_{\theta\in[0,2\pi)}$*. The operator* Γ *is block diagonal with rank* 14 *– it comprises two* 3*-dimensional blocks associated with* $\mathcal{P}_1$ *and* $\mathcal{P}_2$*, and a* 8*-dimensional block associated with* $\mathcal{P}_3$*. We denote the block operators associated with* $\mathcal{P}_i$ *by* Γ_i*. From the block diagonal structure of* Γ*, we conclude that the distributions of* $\mathcal{P}_1$*,* $\mathcal{P}_2$*, and* $\mathcal{P}_3$ *are asymptotically independent. Moreover, the deviations of* $\mathcal{P}_1$ *and* $\mathcal{P}_2$ *from* $\{(0,0,1)'\}$ *and* $\{(0,0,-1)'\}$ *are asymptotically normal with inverse covariance specified by* Γ_1 *and* Γ_2 *respectively.*

We explain the behavior of $\mathcal{P}_3$ in further detail. The operator Γ_3 is the sum of an operator $\Gamma_{3,xy}$ with rank 5 describing the variation of $\mathcal{P}_3$ in the xy-plane, and another operator $\Gamma_{3,z}$ with rank 3 describing the variation of $\mathcal{P}_3$ in the direction of the z-axis. The operator $\Gamma_{3,xy}$, when restricted to the appropriate subspace and suitably scaled, is equal to the operator we encountered in the previous example in the setting where $\mathcal{K}^\star$ is the ℓ_2-ball in $\mathbb{R}^2$, and hence the description of the behavior of $\mathcal{P}_3$ in the xy-plane follows from our prior discussion. The operator $\Gamma_{3,z}$ comprises a single mode representing oscillations of $\mathcal{P}_3$ in the z direction (see subfigure (b) in Figure 3.4), and two modes representing "wobbling" of $\mathcal{P}_3$ with respect to the xy-plane (see subfigures (c) and (d) in Figure 3.4).



(a) Extreme points of $\mathcal{K}^\star$ (b) $(0,0,1)'$ (c) $(0,0,\cos(\theta))'$ (d) $(0,0,\sin(\theta))'$

Figure 3.4: Estimating $\mathcal{K}^\star$ the spectral norm ball in $\mathbb{S}^2$ as the projection of the set C (3.9). The extremal points of the estimator $\hat{\mathcal{K}}_n(C)$ comprise a connected component that is isomorphic to $\mathcal{S}^1$ (see the above accompanying discussion), and the above figure describes the possible modes of oscillations. There are 8 modes altogether – 5 of which occurs in the xy-plane and are described in Figure 3.3, and the remaining 3 are shown in (b),(c), and (d).

Challenges of extending Theorem 3.3.5 to general affine slices. In the introduction, we noted that a more general regression problem leading to broader classes of convex sets C is one where we optimize over affine slices of the nonnegative orthant in $\mathbb{R}^q$ or the cone of positive semidefinite matrices in $\mathbb{S}^p$. We briefly outline the challenges that are involved in obtaining such a result. Suppose we let $\mathcal{P} \subset \mathbb{R}^q$ denote the choice of such a cone. Then a general slice of the cone $\mathcal{P}$ is specified by:

$$B\begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} = 0, \quad B \in L(\mathbb{R}^{q+1}, \mathbb{R}^s),$$

in which case, our regression problem can be generalized to the following:

$$\begin{aligned} \underset{A \in L(\mathbb{R}^q, \mathbb{R}^d), B \in L(\mathbb{R}^{q+1}, \mathbb{R}^r)}{\operatorname{argmin}} \quad & \mathbb{E}_{P_{\mathcal{K}^\star}}(y - h_C(\mathbf{u}))^2 \\ \text{s.t.} \quad & C = \left\{ \mathbf{y} : \mathbf{y} = A\mathbf{x}, B\begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} = 0, \mathbf{x} \in \mathcal{P} \right\}. \end{aligned} \tag{3.10}$$

In principle, one could consider establishing a CLT type of result for the minimizers of (3.10) jointly over A and B; the difficulty in doing so is that an appropriate choice of the dimension s is typically unknown. One perspective is to establish a result for $s = q$, as q dimensions suffices to express all possible affine slices – such an approach is problematic because it leads to degenerate optimal solutions that contain zero singular values. Our earlier description of an instance where $\Phi_C(\cdot, P_{\mathcal{K}^\star})$ is not twice differentiable at the global minimizer indicates the type of behavior that may arise when the minimizers in $\mathcal{M}_{\mathcal{K}^\star,C}$ contain degeneracies. An alternative perspective is to seek a result for a fixed choice of s, where s is chosen to be sufficiently small to avoid degeneracies. In such a setting, it is necessary to account for equivalences that arise due to automorphisms of $\mathcal{P}$ (these are analogous to the equivalences due to $\mathrm{Aut}(C)$ in Section 3.2.2) – one way of doing so is to fix a canonical choice of linear map B, in which case the analysis reduces to an optimization over projection maps, which is precisely the main setting we consider in this paper.

3.3.2.1 Proof of Proposition 3.3.4

Given a function $f(\cdot)$, the graph of f is defined to be the set $\{(\mathbf{x}, s) : 0 \leq s \leq f(\mathbf{x})\} \cup \{(\mathbf{x}, s) : f(\mathbf{x}) \leq s \leq 0\}$. The proof of Proposition 3.3.4 relies on showing that the collections of sets arising as graphs of the following satisfy an appropriate complexity bound and for a radius $r > 0$:

$$\mathfrak{F} := \left\{\lambda_C((\mathbf{u}, y), \hat{A}, D)\right\}_{D \in r\mathcal{B}}, \quad \text{where} \quad \mathcal{B} = \mathcal{B}_{\|\cdot\|_{\mathcal{L}^2(P_{\mathcal{K}^\star})}}(0) \subset L(\mathbb{R}^q, \mathbb{R}^d).$$

Here, $\lambda_C((\mathbf{u}, y), \hat{A}, D)$ is to be viewed as a function in the arguments $(\mathbf{u}, y)$, while D indexes the collection. We do so by appealing to the notion of a *Vapnik-Chervonenkis* class.[3]

Definition 3.3.9 (Vapnik-Chervonenkis (VC) Class) *Let C be a set and $\mathfrak{F}$ be a collection of subsets of C. A finite set $\mathcal{D}$ is said to be* shattered *by $\mathfrak{F}$ if for every $\mathcal{A} \subset \mathcal{D}$ there exists $\mathcal{G} \in \mathfrak{F}$ such that $\mathcal{A} = \mathcal{G} \cap \mathcal{D}$. The collection $\mathfrak{F}$ is said to be* VC class *if there is a finite k such that all sets with cardinality greater than k cannot be shattered by $\mathfrak{F}$.*

[3] It is perhaps more customary, especially in other texts, to show that the *epigraphs* rather than the graphs of a collection of functions is VC. It turns out that the choice is immaterial, as the graphs being VC implies that the epigraphs are also VC, and vice versa. As the proof of Proposition 3.3.4 relies on results in [113], we retain the formalism developed in the same text.

VC classes feature prominently in statistical learning theory; in particular, it provides a framework for describing conditions under which generalization of a learning algorithm is possible. The crux of Proposition 3.3.4 is a result by Stengle and Yukich [136] showing that a large collection of sets that admit semialgebraic descriptions are VC.

Proof of Proposition 3.3.4. The proof relies on computing an entropy bound for the graphs of functions in $\mathfrak{F}$. Define the classes of functions $\mathfrak{F}_1 := \{((h_C((A+D)'\mathbf{u}) - y)^2 - (h_C(A'\mathbf{u}) - y)^2)/\|D\|\}_{D\in r\mathcal{B}}$, and $\mathfrak{F}_2 := \{\langle\nabla(\cdot, A), D/\|D\|\rangle\}_{D\in r\mathcal{B}}$. Note that every element in $\mathfrak{F}$ is expressible as a sum of elements in $\mathfrak{F}_1$ and $\mathfrak{F}_2$.

First we compute entropy bounds for $\mathfrak{F}_1$ and $\mathfrak{F}_2$. We have $p(\mathbf{u}, y, s, D, \mathbf{e}_1, \mathbf{e}_2, s) := \langle A'\mathbf{u} - yI, \mathbf{e}_1\mathbf{e}_1'\rangle^2 - \langle (A+D)'\mathbf{u} - yI, \mathbf{e}_2\mathbf{e}_2'\rangle^2 - s$, where $\mathbf{e}_i \in \mathbb{R}^q$, $i \in \{1,2\}$, is a polynomial in $(\mathbf{u}, y, s, D, \mathbf{e}_1, \mathbf{e}_2, s)$. Hence by Theorem 1 of [136], the collection of sets given by

$$\left\{\left\{(\mathbf{u}, y, s) : \sup_{\|\mathbf{e}_1\|\le 1}\inf_{\|\mathbf{e}_2\|\le 1} p(\mathbf{u}, y, s, D, \mathbf{e}_1, \mathbf{e}_2, s) \ge 0\right\} : D \in r\mathcal{B}\right\}, \tag{3.11}$$

forms a VC class. The collection of sets $\{\{(\mathbf{u}, y, s) : s \ge 0\} : D \in r\mathcal{B}\}$ is also a VC class. Since the intersection of VC classes is a VC class, the collection of sets $\{\{(\mathbf{u}, y, s) : ((h_C((A+D)'\mathbf{u}) - y)^2 - (h_C(A'\mathbf{u}) - y)^2)/\|D\| \ge s \ge 0\} : D \in r\mathcal{B}\}$ is a VC class. A similar sequence of arguments shows that the collection of sets $\{\{(\mathbf{u}, y, s) : ((h_C((A+D)'\mathbf{u}) - y)^2 - (h_C(A'\mathbf{u}) - y)^2)/\|D\| \le s \le 0\} : D \in r\mathcal{B}\}$ is also a VC class. As the union of VC classes is a VC class, we conclude that the graphs of functions in $\mathfrak{F}_1$ form a VC class. Next the set $\mathfrak{F}_2$ is a finite dimensional collection of functions, with dimension equal to that of the linear map D. Hence by Lemma 9.6 of [88] the graphs of functions in $\mathfrak{F}_2$ is a VC class.[4]

Second by noting that $D \in r\mathcal{B}$ is bounded and by applying some elementary computations, the function classes $\mathfrak{F}_1$ and $\mathfrak{F}_2$ have a positive envelop of the form $c(1 + |y|)$, where c is some constant depending only on A.[5]

Third we apply the entropy bound to conclude the result. Our approach is a standard procedure in the literature, and we follow a sequence of arguments that is similar to

[4]The definition of a graph in [88] differs from the definition of a graph in [113]; for consistency we stick with the definition in [113]. One can apply a sequence of operations analogous to what we did for $\mathfrak{F}_1$ to show that the graphs of $\mathfrak{F}_2$ is a VC class.

[5]A function $\tilde{f}$ is a positive envelop for a class of functions $\tilde{\mathfrak{F}}$ if $\tilde{f} \in \mathcal{L}^2(P_{\mathcal{K}^\star})$, and $|f| \le \tilde{f}$ for all $f \in \tilde{\mathfrak{F}}$.

the proof in Example 19 of Chapter VII of [113]. For completeness, we present the necessary arguments.

Let Q be any probability distribution over the variables $(\mathbf{u}, y)$. In addition, let $\mu(\epsilon, Q, \mathfrak{F})$ denote the minimum m such that there exists an ϵ-cover of the form $\{f_i\}_{i=1}^m$ for the collection $\mathfrak{F}$ in the $\mathcal{L}^2(Q)$-metric (the functions f_i are not required to be in $\mathfrak{F}$). Define $\tau(\delta, Q, \mathfrak{F}) := \int_0^\delta \sqrt{\log \mu(t, Q, \mathfrak{F})} dt$. By Lemma II:36 of [113], there exists constants $c_i, d_i, i \in \{1, 2\}$, such that

$$\mu(\delta \sqrt{Q\{f^2\}}, Q, \mathfrak{F}_i) \leq c_i \delta^{-d_i}, \quad i \in \{1, 2\}, \tag{3.12}$$

for all $0 < \delta \leq 1$, all probability distributions Q over the variables $(\mathbf{u}, y)$, and any choice of positive envelop function f. By the AM-GM inequality we have

$$\mu((\delta/2)\sqrt{Q\{f^2\}}, Q, \mathfrak{F}_1) + \mu((\delta/2)\sqrt{Q\{f^2\}}, Q, \mathfrak{F}_2) \leq \mu(\delta \sqrt{Q\{f^2\}}, Q, \mathfrak{F}),$$

and therefore there also exists constants c, d such that $\mu(\delta \sqrt{Q\{f^2\}}, Q, \mathfrak{F}) \leq c\delta^{-d}$. In particular, by choosing $Q = P_{n,\mathcal{K}^\star}$ and $f = c(1 + |y|)$, it follows that

$$\mu\left(\delta \sqrt{\mathbb{E}_{P_{n,\mathcal{K}^\star}}\{c(1 + |y|)\}^2}, P_{n,\mathcal{K}^\star}, \mathfrak{F}\right) \leq c\delta^{-d}, \quad n \geq 0. \tag{3.13}$$

Subsequently, by applying the bound in (3.13) to Lemma 15 of Chapter VII in [113], we conclude that there exists a $\delta > 0$ for which

$$\limsup_n \mathbb{P}_{\mathcal{K}^\star}\left\{ \sup_{\|\lambda_C(\cdot,\hat{A},D_1) - \lambda_C(\cdot,\hat{A},D_2)\|_{\mathcal{L}^2(P_{\mathcal{K}^\star})} \leq \delta} \left| \mathbb{E}_{E_n}\left\{\lambda_C(\cdot, \hat{A}, D_1) - \lambda_C(\cdot, \hat{A}, D_2)\right\}\right| > \eta \right\} < \epsilon.$$

This implies the desired result. □

Remark. *The only argument in the proof of Proposition 3.3.4 that pertains to C being the free spectrahedron is the existence of a polynomial $p(\cdot)$ such that sets in the collection $\mathfrak{F}$ can be expressed in the form of* (3.11)*. For many families of convex sets C that admit semialgebraic descriptions, such a polynomial $p(\cdot)$ exists, in which case the sequence of stochastic processes $\{\{\mathbb{E}_{E_n}\{\lambda_C(\cdot, \hat{A}, D)\} : D \in L(\mathbb{R}^q, \mathbb{R}^d)\}\}_{n=1}^\infty$ too, is also stochastically equicontinuous.*

3.3.3 Preservation of Facial Geometry

Our third result describes the manner in which estimators computed using our method respect the facial geometry of the underlying set.

We begin our discussion with a motivating numerical experiment in which we reconstruct the ℓ_1 ball in $\mathbb{R}^3$ from 200 noisy support function evaluations. More precisely, we apply our method with the simplex in $\mathbb{R}^6$ as the choice of lifting set, and we compare the resulting estimate with the LSE in Figure 3.5. Our results show a one-to-one correspondence between the faces of the reconstruction obtained using our method (second subfigure from the left) with the faces of the ℓ_1 ball (leftmost subfigure); in contrast, we do not observe an analogous correspondence between the faces of the LSE (rightmost subfigure) and the faces of the ℓ_1 ball.

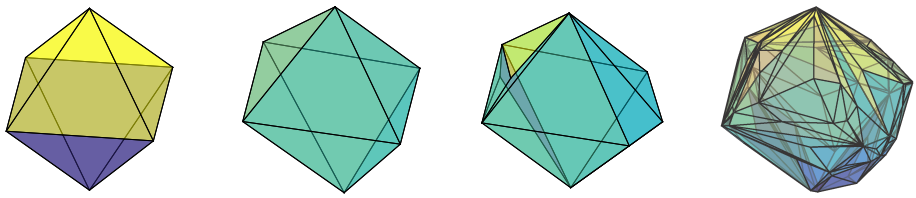


Figure 3.5: Reconstructions of the unit ℓ_1-ball (left) in $\mathbb{R}^3$ from 200 noisy support function measurements using our method with $C = \Delta^6$ (second from left), and with $C = \Delta^{12}$ (third from left). The LSE is the rightmost figure.

The statement of our result requires us to formalize the notion by which a sequence of estimators 'respects the facial geometry' of an underlying set. The precise manner in which we do so is via the existence of an invertible affine transformation between the faces of the underlying set and the faces of the reconstruction. For analytical tractability, our result focuses on exposed faces – these are faces that are expressible as the intersection of the underlying set with a half-space.

Definition 3.3.10 *Let $\{\mathcal{K}_n\}_{n=1}^{\infty} \subset \mathbb{R}^d$ be a sequence of compact convex sets converging to some $\mathcal{K} \subset \mathbb{R}^d$. Let $\mathcal{F} \subset \mathcal{K}$ be an exposed face. We say that $\mathcal{F}$ is* preserved *by the sequence $\{\mathcal{K}_n\}_{n=1}^{\infty}$ if there is a sequence $\{\mathcal{F}_n\}_{n=n_0}^{\infty}$, $\mathcal{F}_n \subseteq \mathcal{K}_n$, satisfying*

1. *$\mathcal{F}_n \to \mathcal{F}$.*
2. *$\mathcal{F}_n$ are exposed faces of $\mathcal{K}_n$.*
3. *There is an invertible affine transformation B_n such that $\mathcal{F} = B_n(\mathcal{F}_n)$ and $\mathcal{F}_n = B_n^{-1}(\mathcal{F})$.*

Our main result in this subsection provides sufficient conditions under which an exposed face $\mathcal{F}^\star \subset \mathcal{K}^\star$ is preserved by a sequence of estimators $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^{\infty}$ obtained using our method. It operates under two sets of conditions:

1. The first set of conditions concern the choice of lifting set $C \subset \mathbb{R}^q$ in relation to the underlying set $\mathcal{K}^\star \subset \mathbb{R}^d$; more precisely, we require $\mathcal{K}^\star$ to be representable as a projection of C, and that the set of minimizers $\mathcal{M}_{\mathcal{K}^\star,C}$ form a single orbit of a linear map $A^\star$ under the action of the automorphism group of C, i.e., $\mathcal{M}_{\mathcal{K}^\star,C} = \mathfrak{D}_C(A^\star)$ for some $A^\star \in L(\mathbb{R}^q, \mathbb{R}^d)$. To provide some intuition as to what these conditions impose, in the simplest setting where the set $\mathcal{K}^\star$ is polyhedral and we choose C from among the family of simplices $\{\Delta^q\}_{q=1}^\infty$, these conditions can be easily shown to be equivalent to selecting C as the simplex in dimensions equal to the number of extreme points of $\mathcal{K}^\star$. It is clear that an inadequate lifting dimension would not provide sufficient expressive power in our framework to express $\mathcal{K}^\star$; on the other hand, the manner in which our method fails to recover the correct facial geometry when the lifting dimension is excessive is quite different, and it can be attributed to the inclusion of spurious extreme points. To give an illustration, we revisit the previous numerical experiment in which we estimate the ℓ_1 ball in $\mathbb{R}^3$, and apply our method with Δ^{12} as the choice of our lifting set as opposed to Δ^6. Our reconstruction in Figure 3.5 (see the third subfigure from the left) preserves a subset of the faces of the ℓ_1 ball, and breaks the other faces into smaller simplicial faces. Although our reconstruction fails to capture the correct facial geometry, it nevertheless bears a stronger resemblance (qualitatively speaking) to the underlying set as compared to the LSE.

2. Assuming that the first set of requirements are satisfied, the second set of conditions concern the face $\mathcal{F}^\star$ and its relation to the lifting set C. These conditions can be viewed as a characterization of instances under which $\mathcal{F}^\star$ continues to be an exposed face under small perturbations (of an appropriate type). More precisely, let $\mathcal{G} \subset C$ be the pre-image of $\mathcal{F}^\star$ under the projection map $A^\star$. First, we require $A^\star$ to be injective over the affine hull of $\mathcal{G}$. In the simplest setting where $\mathcal{K}^\star$ is polyhedral, and we specify C to be a simplex in dimensions equal to the number of extreme points of $\mathcal{K}^\star$, our requirement concerning the injectivity of $A^\star$ implies that $\mathcal{F}^\star$ must be simplicial. It is instructive to examine what happens when $\mathcal{F}^\star$ is *not* simplicial. In Figure 3.6 we apply our method to an instance where the underlying set is the ℓ_∞ ball in $\mathbb{R}^3$, and we observe that our resulting estimate breaks every face of the ℓ_∞ ball into smaller simplicial faces. Second, we require the dimension of the linear span of the *normal cone* of $\mathcal{G}$ with respect to C to be sufficiently large; here, recall that for any convex $\mathcal{K}$ and any subset $\mathcal{F} \subseteq \mathcal{K}$, the *normal cone* of $\mathcal{F}$

with respect to $\mathcal{K}$ is defined as the following:

$$\mathfrak{R}_{\mathcal{K}}(\mathcal{F}) := \{\mathbf{x} : \langle \mathbf{y}-\mathbf{z}, \mathbf{x}\rangle \leq 0 \text{ for all } \mathbf{y} \in \mathcal{K}, \mathbf{z} \in \mathcal{F}, \text{ and } \langle \mathbf{y}-\mathbf{z}, \mathbf{x}\rangle = 0 \text{ for all } \mathbf{y}, \mathbf{z} \in \mathcal{F}\}.$$

Our stipulation on the size of $\dim(\mathrm{Span}(\mathfrak{R}_C(\mathcal{G})))$ arises because we need to consider how the extreme points of C in the neighborhood of the pre-image $\mathcal{G}$ affect the facial geometry of our estimate of $\mathcal{F}^\star$. In the simplest instance where $\mathcal{K}^\star$ is polyhedral and we choose C to be the simplex, our condition is trivially satisfied as the extreme points of C are isolated (we discuss this point in detail subsequently); the situation becomes more delicate when $\mathcal{K}^\star$ and C are non-polyhedral. We illustrate what may happen if $\dim(\mathrm{Span}(\mathfrak{R}_C(\mathcal{G})))$ is not sufficiently large with a numerical experiment in which we estimate the *Race Track* in $\mathbb{R}^2$ from 200 noisy support function measurements:

$$\text{Race Track} := \mathrm{conv}(\{(x,y)' : \|(x,y)'-(-1,0)'\|_2 \leq 1 \text{ or } \|(x,y)'-(1,0)'\|_2 \leq 1\}).$$

More precisely, we consider the recovery of the two horizontal faces of the Race Track. In Figure 3.7 we show a reconstruction as the projection of O^4, and we observe that the estimates of the straight edges of the Race Track are curved.

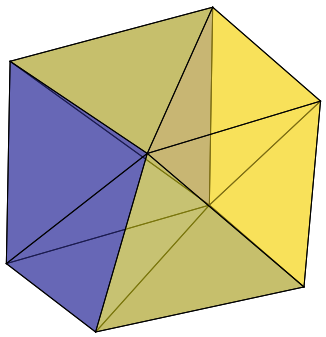


Figure 3.6: Reconstruction of the unit ℓ_∞ ball in $\mathbb{R}^3$ from 75 noisy support function measurements using our method. The choice of lifting set is $C = \Delta^8$.

Theorem 3.3.11 *Suppose that $\mathcal{K}^\star \subseteq \mathbb{R}^d$ is a compact convex set with non-empty interior. Let $C \subset \mathbb{R}^q$ be a compact convex set such that* $\mathrm{Span}(C) \cong \mathbb{R}^q$. *Suppose that there is a linear map $A^\star \in L(\mathbb{R}^q, \mathbb{R}^d)$ such that $\mathcal{K}^\star = A^\star(C)$, and $\mathcal{M}_{\mathcal{K}^\star, C} = \mathfrak{D}_C(A^\star)$. Let $\{\hat{A}_n\}_{n=1}^\infty$, $\hat{A}_n \in \mathrm{argmin}_A \Phi_C(A, P_{n,\mathcal{K}^\star})$, be a sequence of minimizers of the empirical loss function, and let $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^\infty$, $\hat{\mathcal{K}}_n(C) = \hat{A}_n(C)$, be the corresponding sequence of estimators of $\mathcal{K}^\star$. Given an exposed face $\mathcal{F}^\star \subset \mathcal{K}^\star$, and let $\mathcal{G} = \{\mathbf{x} : A^\star\mathbf{x} \in \mathcal{F}^\star\} \cap C$ be its pre-image. If*

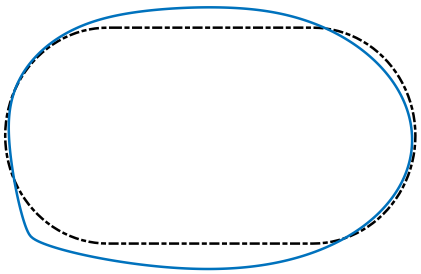


Figure 3.7: Reconstruction of the *Race Track* from 200 noisy support function measurements using our method. The choice of lifting set is the free spectrahedron $\mathcal{O}^4$.

1. *the linear map* $A^\star$ *is injective when restricted to* $\mathrm{aff}(\mathcal{G})$*, and*
2. $\dim(\mathrm{Span}(\mathfrak{R}_C(\mathcal{G}))) > q - \mathrm{rank}(A^\star)$*,*

then $\mathcal{F}^\star$ *is preserved by the sequence* $\{\hat{\mathcal{K}}_n(C)\}_{n=1}^\infty$.

Proof of Theorem 3.3.11. As we noted in the above, define $\tilde{A}_n \in \mathrm{argmin}_{A\in\mathfrak{D}_C(\hat{A}_n)}\|\hat{A}_n - A\|_F$, and denote $\mathcal{F}_n = \tilde{A}_n(\mathcal{G})$.

[$\mathcal{F}_n \to \mathcal{F}$]: Since $\mathcal{M}_{\mathcal{K}^\star,C} = \mathfrak{D}_C(A^\star)$, it follows from Theorem 3.3.1 that $\tilde{A}_n \to A^\star$, from which we have $\mathcal{F}_n \to \mathcal{F}^\star$.

[$\mathcal{F}_n$ are faces of $\mathcal{K}_n$]: Since $\mathcal{F}^\star$ is an exposed face of $\mathcal{K}^\star$, there exists $\mathbf{y} \in \mathbb{R}^d$ and $c \in \mathbb{R}$ such that $\langle \mathbf{y}, \mathbf{x}\rangle = c$ for all $\mathbf{x} \in \mathcal{F}^\star$, and $\langle \mathbf{y}, \mathbf{x}\rangle > c$ for all $\mathbf{x} \in \mathcal{K}^\star \backslash \mathcal{F}^\star$. This implies that $\langle A^{\star\prime}\mathbf{y}, \tilde{\mathbf{x}}\rangle = c$ for all $\tilde{\mathbf{x}} \in \mathcal{G}$, and $\langle A^{\star\prime}\mathbf{y}, \tilde{\mathbf{x}}\rangle > c$ for all $\tilde{\mathbf{x}} \in C\backslash\mathcal{G}$. In particular, it implies that the row space of $A^\star$ intersects the relative interior of $\mathfrak{R}_C(\mathcal{G})$ in the direction $A^\star\mathbf{y}$.

By combining the earlier conclusion that $\tilde{A}_n \to A^\star$ a.s., and that $\dim(\mathrm{Span}(\mathfrak{R}_C(\mathcal{G})))+\mathrm{rank}(A^\star) > q$, we conclude that the row spaces of the maps $\tilde{A}_n$ eventually intersect the relative interior of $\mathfrak{R}_C(\mathcal{G})$ a.s.. That is to say, there is exists an integer n_0 and sequences $\{\mathbf{y}_n\}_{n=n_0}^\infty \subset \mathbb{R}^d$, $\{c_n\}_{n=n_0}^\infty \subset \mathbb{R}$ such that $\langle \mathbf{y}_n, \mathbf{x}\rangle = c_n$ for all $\mathbf{x} \in \mathcal{F}_n$, and $\langle \mathbf{y}_n, \mathbf{x}\rangle > c$ for all $\mathbf{x} \in \hat{\mathcal{K}}_n(C)\backslash\mathcal{F}_n$, $n \geq n_0$, a.s.. In other words, the sets $\mathcal{F}_n$ are exposed faces of $\hat{\mathcal{K}}_n(C)$ eventually a.s..

[One-to-one affine correspondence]: To establish a one-to-one affine correspondence between $\mathcal{F}_n$ and $\mathcal{F}$ we need to treat the case where $0 \in \mathrm{aff}(\mathcal{G})$ and the case where $0 \notin \mathrm{aff}(\mathcal{G})$ separately.

First suppose that $0 \in \mathrm{aff}(\mathcal{G})$. Let $\mathfrak{H}_\mathcal{F} = \mathrm{aff}(\mathcal{F})$ and $\mathfrak{H}_\mathcal{G} = \mathrm{aff}(\mathcal{G})$. Since $0 \in \mathfrak{H}_\mathcal{G}$, it follows that $\mathfrak{H}_\mathcal{F}$ and $\mathfrak{H}_\mathcal{G}$ are subspaces. Moreover given that $A^\star$ is injective restricted

to $\mathfrak{H}_{\mathcal{G}} = \text{aff}(\mathcal{G})$, it follows that $\mathfrak{H}_{\mathcal{F}}$ and $\mathfrak{H}_{\mathcal{G}}$ have equal dimensions. Hence the map T defined as the restriction of $A^\star$ onto $L(\mathfrak{H}_{\mathcal{G}}, \mathfrak{H}_{\mathcal{F}})$ is square and invertible. Next let $\mathfrak{H}_{\mathcal{F}_n} = \text{aff}(\mathcal{F}_n)$, and let T_n denote the restriction of $\tilde{A}_n$ to $L(\mathfrak{H}_{\mathcal{G}}, \mathfrak{H}_{\mathcal{F}_n})$. Given that $\tilde{A}_n \to A^\star$, the maps $\{T_n\}_{n=1}^{\infty}$ are also square and invertible eventually a.s.. It follows that one can define a linear map $B_n \in L(\mathbb{R}^d, \mathbb{R}^d)$ that coincides with $T \circ T_n^{-1}$ restricted to $L(\mathfrak{H}_{\mathcal{F}_n}, \mathfrak{H}_{\mathcal{F}})$, is permitted to be any square invertible map on $L(\mathfrak{H}_{\mathcal{F}_n}^{\perp}, \mathfrak{H}_{\mathcal{F}}^{\perp})$, and is zero everywhere else. Notice that B_n is invertible by construction. It straightforward to check that $\mathcal{F} = B_n(\mathcal{F}_n)$ and $\mathcal{F}_n = B_n^{-1}(\mathcal{F})$.

Next suppose that $0 \notin \text{aff}(\mathcal{G})$. The treatment in this case is largely similar as in the previous case. Let $\mathfrak{H}_{\mathcal{F}}$ be the smallest subspace containing $\{(\mathbf{x}, 1) : \mathbf{x} \in \mathcal{F}\} \subseteq \mathbb{R}^{d+1}$, where the set $\mathcal{F}$ is embedded in the first d coordinates. Let $\mathfrak{H}_{\mathcal{F}_n}$ be similarly defined. Let $\mathfrak{H}_{\mathcal{G}} = \text{aff}(\mathcal{G} \cup \{0\})$ – note that this defines a subspace. Since $0 \notin \text{aff}(\mathcal{G})$, there is a nonzero $\mathbf{z} \in \mathbb{R}^q$ such that $\langle \mathbf{z}, \mathbf{x} \rangle = 1$ for all $\mathbf{x} \in \mathcal{G}$ (i.e. there exists a hyperplane containing $\mathcal{G}$). Define the linear map $T \in L(\mathfrak{H}_{\mathcal{G}}, \mathfrak{H}_{\mathcal{F}})$ as

$$T = P_{\mathfrak{H}_{\mathcal{F}}} \left[\left(\begin{array}{c} A^\star \\ \mathbf{z}' \end{array} \right) \Bigg|_{\mathfrak{H}_{\mathcal{G}}} \right]$$

where $P_{\mathfrak{H}_{\mathcal{F}}}$ is the restriction operator onto the subspace $\mathfrak{H}_{\mathcal{F}}$. Since $A^\star$ is injective on $\mathcal{G}$, it follows that $\mathfrak{H}_{\mathcal{F}}$ and $\mathfrak{H}_{\mathcal{G}}$ have the same dimensions, and that T is square and invertible. One can define a square invertible map T_n analogously. The remainder of the proof proceeds in a similar fashion to the previous case, and we omit the details. Here, note that a linear invertible map operating on the lifted space $\mathbb{R}^{d+1}$ defines an affine linear invertible map in the embedded space $\mathbb{R}^d$. □

Remark. *Suppose that $\mathcal{K}^\star$ is a full bodied polytope with q extreme points and we choose $C = \Delta^q$. Then Theorem 3.3.11 implies that all* proper simplicial *faces of $\mathcal{K}^\star$ are preserved. It is easy to see that there is a linear map $A^\star$ such that $\mathcal{K}^\star = A^\star(\Delta^q)$, and that $\mathcal{M}_{\mathcal{K}^\star, \Delta^q} = \mathfrak{D}_{\Delta^q}(A^\star)$. Let $\mathcal{F}^\star \subseteq \mathcal{K}^\star$ be any face (note that all faces of a polytope are always exposed), and let $\mathcal{G}$ be its pre-image in Δ^q. Recall that the pre-image of an exposed face is also an exposed face (see for instance the first part of the proof of Theorem 3.3.11), and hence $\mathcal{G}$ is of the form $\{\Pi \mathbf{x} : \mathbf{x} \geq 0, \langle \mathbf{1}, \mathbf{x} \rangle = 1, \mathbf{x}_{s+1} = \ldots \mathbf{x}_q = 0\}$, for some $\Pi \in \text{Aut}(\Delta^q)$, and some $s \leq q$. We proceed to interpret the remaining requirements. First if $A^\star$ is injective on* aff($\mathcal{G}$)*, then the image of $\mathcal{G}$ under $A^\star$ is isomorphic to $\mathcal{G}$; i.e., $\mathcal{F}^\star$ is simplicial. Second the normal cone $\mathfrak{N}_{\Delta^q}(\mathcal{G})$ is given by $\{\Pi \mathbf{z} : \mathbf{z} \leq 0, \mathbf{z}_1 = \ldots \mathbf{z}_s = 0\}$, and hence the requirement* $\dim(\text{aff}(\mathfrak{N}_{\Delta^q}(\mathcal{G}))) > q - \text{rank}(A^\star)$ *holds precisely when $s < d$; i.e., the face $\mathcal{F}^\star$ is* proper.

Remark. *Suppose that $\mathcal{K}^\star$ is a set that is expressible as the projection of the free spectrahedron $\mathcal{C} = \mathcal{O}^p$ via the linear map $A^\star$, and that we also have $\mathcal{M}_{\mathcal{K}^\star,\mathcal{C}} = \mathfrak{D}_{\mathcal{C}}(A^\star)$. Let $\mathcal{F}^\star$ be an exposed face, and let $\mathcal{G}$ be the pre-image $\mathcal{O}^p$. Then $\mathcal{G}$ must be a face of $\mathcal{O}^p$, and hence is of the form*

$$\mathcal{G} = \left\{ UDU' \ : \ D = \begin{pmatrix} D_{11} & 0 \\ 0 & 0 \end{pmatrix},\ D_{11} \in \mathcal{O}^r \right\},$$

for some $U \in O(p)$, and some $r \leq p$. Note that

$$\mathfrak{R}_{\mathcal{O}^p}(\mathcal{G}) = \left\{ UDU' \ : \ D = \begin{pmatrix} 0 & 0 \\ 0 & -D_{22} \end{pmatrix},\ D_{22} \in \mathbb{S}^{p-r},\ D_{22} \succ 0 \right\},$$

and hence the requirement that $\dim(\mathrm{aff}(\mathfrak{R}_{\mathcal{O}^p}(\mathcal{G}))) > \binom{p+1}{2} - \mathrm{rank}(A^\star)$ *holds precisely when* $d > pr - \binom{r-1}{2}$.

Remark. *Consider our earlier example where we computed an estimate of the Race Track as the projection of $\mathcal{O}^4$. One can check that we may choose $A^\star$ to be the following linear map*

$$A^\star(X) = \begin{pmatrix} \langle A_1, X \rangle \\ \langle A_2, X \rangle \end{pmatrix}, \quad A_1 = \begin{pmatrix} -1 & 1 & & \\ 1 & -1 & & \\ & & 1 & 1 \\ & & 1 & 1 \end{pmatrix} \quad A_2 = \begin{pmatrix} 1 & & & \\ & -1 & & \\ & & 1 & \\ & & & -1 \end{pmatrix}.$$

It is clear that $\mathrm{rank}(A^\star) = 2$. *Let $\mathcal{F}^\star$ be the face connecting $(-1,0)'$ and $(1,0)'$, and let $\mathcal{G}_{\mathcal{O}^4}$ be the pre-image of $\mathcal{F}^\star$ in $\mathcal{O}^4$. One can check that*

$$\mathcal{G}_{\mathcal{O}^4} = \left\{ \begin{pmatrix} x & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & y & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} : x, y \geq 0, x + y \leq 1 \right\}, \mathfrak{R}_{\mathcal{O}^4}(\mathcal{G}_{\mathcal{O}^4}) \left\{ Z : Z = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & * & 0 & * \\ 0 & 0 & 0 & 0 \\ 0 & * & 0 & * \end{pmatrix}, Z \preceq 0 \right\}.$$

It follows that $\dim(\mathrm{aff}(\mathfrak{R}_{\mathcal{O}^4}(\mathcal{G}_{\mathcal{O}^4}))) = 3$, *and since the lifting dimension is* 10, *our requirement on* $\dim(\mathrm{aff}(\mathfrak{R}_{\mathcal{O}^4}(\mathcal{G}_{\mathcal{O}^4})))$ *is not satisfied.*

3.4 Algorithms

We describe two procedures for solving the optimization problem (3.2). One can check that the problem can be reformulated as follows:

$$\underset{A \in L(\mathbb{R}^q, \mathbb{R}^d)}{\mathrm{argmin}} \quad \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - h_{\mathcal{C}}(A' \mathbf{u}^{(i)}) \right)^2. \tag{3.14}$$

Our first algorithm is based on gradient descent and the second algorithm is a form of alternating minimization. We highlight the connection between the alternating approach and Lloyd's algorithm for K-means clustering [93]. As described previously, the problem (3.14) is nonconvex as formulated; consequently, the algorithms discussed next are not guaranteed to return a globally optimal solution. However, we demonstrate the effectiveness of these methods with random initialization in numerical experiments in Section 3.5.

3.4.1 Gradient Descent

Our first approach is based on gradient descent. From Proposition 3.2.5 we have that the derivative of the map $A \to \Phi(A, P_n)$ – if it exists – is given by

$$\nabla_A(\Phi(A, P_n)) = \frac{2}{n} \sum_{i=1}^{n} \left(h_C(A'\mathbf{u}^{(i)}) - y^{(i)} \right) \mathbf{u}^{(i)} \otimes \boldsymbol{e}_C(A'\mathbf{u}^{(i)}).$$

Based on the discussion in Section 3.2.2, this derivative exists for generic A when C is either the simplex or the free spectrahedron; in particular, for these two cases the computation of $h_C, \boldsymbol{e}_C$ is given in Section 3.2.2. We summarize the steps in Algorithm 4.

Algorithm 4 Convex Set Regression via Gradient Descent

Input: A collection $\{(\mathbf{u}^{(i)}, y^{(i)})\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ of support function evaluations; a compact convex set $C \subset \mathbb{R}^q$; an initialization $A \in L(\mathbb{R}^q, \mathbb{R}^d)$; a step size $\eta \in (0, 1]$
Algorithm: Repeat until convergence
1. $D \leftarrow \frac{1}{n} \sum_{i=1}^{n} (h_C(A'\mathbf{u}^{(i)}) - y^{(i)}) \mathbf{u}^{(i)} \otimes \boldsymbol{e}_C(A'\mathbf{u}^{(i)})$
2. $A \leftarrow A - \eta D$
Output: Final iterate A

Although the algorithm presented is based on a fixed stepsize, in practice stepsizes may also be chosen via some type of line search.

3.4.2 Alternating Minimization

The second algorithm is based on iteratively computing solutions to linearizations of the first-order optimality conditions of (3.14). Observing that $h_C(A'\mathbf{u}) = \langle A'\mathbf{u}, \boldsymbol{e}_C(A'\mathbf{u}) \rangle = \langle A, \mathbf{u} \otimes \boldsymbol{e}_C(A'\mathbf{u}) \rangle$, the first-order optimality conditions of (3.14) may be expressed as follows:

$$\begin{aligned} 0 &= \frac{1}{n} \sum_{i=1}^{n} \left(h_C(A'\mathbf{u}^{(i)}) - y^{(i)} \right) \mathbf{u}^{(i)} \otimes \boldsymbol{e}_C(A'\mathbf{u}^{(i)}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\langle A, \mathbf{u}^{(i)} \otimes \boldsymbol{e}_C(A'\mathbf{u}^{(i)}) \rangle - y^{(i)} \right) \mathbf{u}^{(i)} \otimes \boldsymbol{e}_C(A'\mathbf{u}^{(i)}). \end{aligned} \tag{3.15}$$

Ideally, one would like to compute solutions to (4.5) directly; the difficulty in doing so is that the function $\boldsymbol{e}_C(A'\mathbf{u}^{(i)})$ depends non-linearly on A. Our approach to getting around this difficulty is to alternately perform the following operations – fix A and compute each $\boldsymbol{e}_C(A'\mathbf{u}^{(i)})$, and treat the $\boldsymbol{e}_C(A'\mathbf{u}^{(i)})$'s as fixed and update A.

We note that the latter update step may be viewed as the solution of a linear system, and that the success of the procedure we propose is governed by the conditioning of this system. More precisely, suppose $\hat{A}$ is a minimizer of (3.14), and let $\Gamma \in L(L(\mathbb{R}^q, \mathbb{R}^d), L^*(\mathbb{R}^q, \mathbb{R}^d))$ be the operator defined as:

$$\Gamma(A) = \sum_{i=1}^{n} \left\langle A, \mathbf{u}^{(i)} \otimes \boldsymbol{e}_C(\hat{A}'\mathbf{u}^{(i)}) \right\rangle \mathbf{u}^{(i)} \otimes \boldsymbol{e}_C(\hat{A}'\mathbf{u}^{(i)}).$$

The conditioning of the linear system depends on Γ being positive definite; unfortunately, such a condition is not guaranteed to be satisfied, and hence we apply an intermediate Tikhonov regularization step to resolve these issues. We summarize the full procedure in Algorithm 5.

Algorithm 5 Convex Regression via Alternating Minimization

Input: A collection $\{(\mathbf{u}^{(i)}, y^{(i)})\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ of support function evaluations; a compact convex set $C \subset \mathbb{R}^q$; an initialization $A \in L(\mathbb{R}^q, \mathbb{R}^d)$; a choice of regularization parameter $\gamma > 0$

Algorithm: Repeat until convergence

1. $\boldsymbol{e}^{(i)} \leftarrow \boldsymbol{e}_C(A'\mathbf{u}^{(i)})$

2. $V \leftarrow \left(\mathbf{u}^{(1)} \otimes \boldsymbol{e}^{(1)} | \ldots | \mathbf{u}^{(n)} \otimes \boldsymbol{e}^{(n)}\right), \quad Y \leftarrow \left(y^{(1)}, \ldots, y^{(n)}\right)'$

3. $A \leftarrow (V \otimes V + \gamma I)^{-1}(VY + \gamma A)$

Output: Final iterate A

Connection to Lloyd's algorithm. When $C = \Delta^q$, Algorithm 5 is similar to Lloyd's algorithm for K-means clustering [93]. Specifically, Lloyd's algorithm begins with an initialization of q centers, and it alternates between (i) assigning data-points to centers based proximity (keeping the centers fixed), and (ii) updating the location of cluster centers to minimize the squared-loss error. In our context, suppose we express the linear map $A = [\mathbf{a}_1 | \ldots | \mathbf{a}_q] \in \mathbb{R}^{d \times q}$ in terms of its columns. The algorithm begins with an initialization of the q columns, and it alternates between (i) assigning measurement pairs $(\mathbf{u}^{(i)}, y^{(i)})$, $1 \leq i \leq n$, to the respective columns $\{\mathbf{a}_j\}_{1 \leq j \leq q}$ such that the inner product $\langle \mathbf{u}^{(i)}, \mathbf{a}_j \rangle$ is maximized (keeping the columns fixed), and (ii) updating the columns $\{\mathbf{a}_j\}_{1 \leq j \leq q}$ to minimize the squared-loss error.

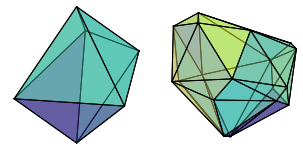
3.5 Numerical Experiments

In this section we describe the results of numerical experiments on fitting convex sets to support function evaluations in which we contrast our framework based on solving (3.2) to previous methods based on solving (3.1). The first few experiments are on synthetically generated data, while the final experiment is on a reconstruction problem with real data obtained from the Computed Tomography (CT) scan of a human lung. For each experiment, we apply both the algorithms described in Section 3.4 with multiple random initializations, and we select the solution that minimizes the least squared error. Specifically, we begin with Algorithm 5, and in instances where the procedure fails to converge after a fixed number of iterations we restart using Algorithm 4. We observe that Algorithm 4 converges in all instances we consider, while Algorithm 5 exhibits much faster convergence compared to Algorithm 4 on those instances in which it converges. The (polyhedral) LSE reconstructions in our experiments are based on the algorithm proposed in [60, Section 4].

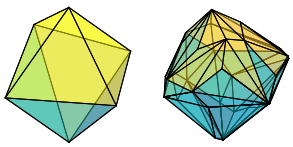
3.5.1 Reconstructing the ℓ_1-ball and the ℓ_2-ball

We consider reconstructing the ℓ_1-ball $\{\mathbf{g} : \|\mathbf{g}\|_1 \leq 1\} \subset \mathbb{R}^3$ and the ℓ_2-ball $\{\mathbf{g} : \|\mathbf{g}\|_2 \leq 1\} \subset \mathbb{R}^3$ from noiseless and noisy support function evaluations based on the model (3.2). In particular, we evaluate the performance of our framework relative to the reconstructions provided by the LSE for $n = 20, 50, 200$ measurements. For both the ℓ_1-ball and the ℓ_2-ball in the respective noisy cases, the measurements are corrupted with additive Gaussian noise of variance $\sigma^2 = 0.1$. The reconstructions based on our framework (3.2) of the ℓ_1-ball employ the choice $C = \Delta^6$, while those of the ℓ_2-ball use $C = O^3$. Figure 3.8 and Figure 3.9 give the results corresponding to the ℓ_1-ball and the ℓ_2-ball, respectively.

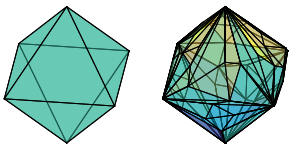
Considering first a setting with noiseless measurements, we observe that our approach gives an exact reconstruction for both the ℓ_1-ball and the ℓ_2-ball. For the ℓ_1-ball this occurs when we have $n = 200$ measurements, while the LSE provides a reconstruction with substantially more complicated facial structure that doesn't reflect that of the ℓ_1-ball. Indeed, the LSE only approaches the ℓ_1-ball in the sense of Hausdorff metric, but despite being the best solution in terms of minimizing the least-squares criterion, the reconstruction offered by this method provides little information about the facial geometry of the ℓ_1-ball. Further, even with $n = 20, 50$ measurements, our reconstructions bear far closer resemblance to the ℓ_1-ball, while the LSE in these cases looks very different from the ℓ_1-ball. For the ℓ_2-ball, our approach provides an exact reconstruction with just $n = 20$ measurements, while



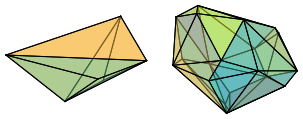
(a) 20 noiseless measurements



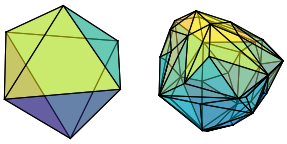
(b) 50 noiseless measurements



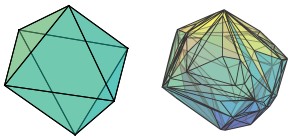
(c) 200 noiseless measurements



(d) 20 noisy measurements



(e) 50 noisy measurements



(f) 200 noisy measurements

Figure 3.8: Reconstruction of the unit ℓ_1-ball in $\mathbb{R}^3$ from noiseless (first row) and noisy (second row) support function measurements. The reconstructions obtained using our method (with $C = \Delta^6$ in (3.2)) are the on the left of every subfigure, while the LSE reconstructions are on the right of every subfigure.

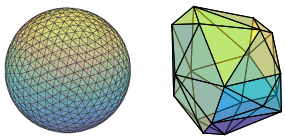
the LSE only begins to resemble the ℓ_2-ball with $n = 200$ measurements (and even then, the reconstruction is a polyhedral approximation).

Turning our attention next to the noisy case, the contrast between the results obtained using our framework and those of the LSE approach is even more stark. For both the ℓ_1-ball and the ℓ_2-ball, the LSE reconstructions bear little resemblance to the underlying convex set, unlike the estimates produced using our method. Notice that the reconstructions of the ℓ_2-ball using our algorithm are not even ellipsoidal when the number of measurements is small (e.g., when $n = 20$), as linear images of the free spectrahedron $\mathcal{O}^3$ may be non-ellipsoidal in general and need not even consist of smooth boundaries. Nonetheless, as the number of measurements available to our algorithm increases, the estimates improve in quality and offer improved reconstructions – with smooth boundaries – of the ℓ_2-ball.

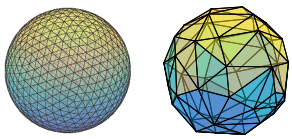
In summary, these synthetic examples demonstrate that our framework is much more effective than the LSE in terms of robustness to noise, accuracy of reconstruction given a small number of measurements, and in settings in which the underlying set is non-polyhedral.

3.5.2 Reconstruction via Linear Images of the Free Spectrahedron

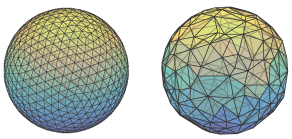
In the next series of synthetic experiments, we consider reconstructions of convex sets with non-smooth boundaries via linear images of the free spectrahedron. In



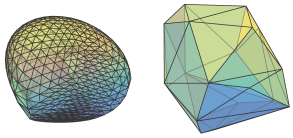
(a) 20 noiseless measurements



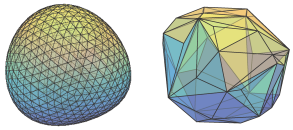
(b) 50 noiseless measurements



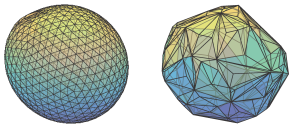
(c) 200 noiseless measurements



(d) 20 noisy measurements



(e) 50 noisy measurements



(f) 200 noisy measurements

Figure 3.9: Reconstruction of the unit ℓ_2-ball in $\mathbb{R}^3$ from noiseless (first row) and noisy (second row) support function measurements. The reconstructions obtained using our method (with $C = O^3$ in (3.2)) are the on the left of every subfigure, while the LSE reconstructions are on the right of every subfigure.

these illustrations, we consider sets in $\mathbb{R}^2$ and in $\mathbb{R}^3$ for which noiseless support function evaluations are obtained and supplied as input to the problem (3.2), with C equal to a free spectrahedron O^q in some larger-dimensional space q. For the examples in $\mathbb{R}^2$, the support function evaluations are obtained at 1000 equally spaced points on the unit circle $\mathcal{S}^1$. For the examples in $\mathbb{R}^3$, the support function evaluations are obtained at 2562 regularly spaced points on the unit sphere $\mathcal{S}^2$ based on an icosphere discretization.

We consider reconstruction of the ℓ_1-ball in $\mathbb{R}^d$. Figure 3.10 shows the output from our algorithm when $d = 2$ for $q \in \{2, 3, 4\}$, and the reconstruction is exact for $q = 4$. Figure 3.11 shows the output from our algorithm when $d = 3$ for $q \in \{3, 4, 5, 6\}$. Interestingly, when $d = 3$ the computed solution for $q = 5$ does not contain any isolated extreme point (i.e., vertices) even though such features are expressible as projections of the free spectrahedron O^5.

As our next illustration, we consider the following projection of O^4:

$$\text{UPillow} = \left\{(x, y, z)' : X \in O^4, X_{12} = X_{21} = x, X_{23} = X_{32} = y, X_{34} = X_{43} = z\right\} \subset \mathbb{R}^3. \tag{3.16}$$

We term this convex set as the 'uncomfortable pillow' and it contains both smooth and non-smooth components in its boundary. Figure 3.12 shows the reconstruction of UPillow as linear images of O^3 and O^4 computed using our algorithm. The reconstruction based on O^4 is exact, while the reconstruction based on O^3 smooths

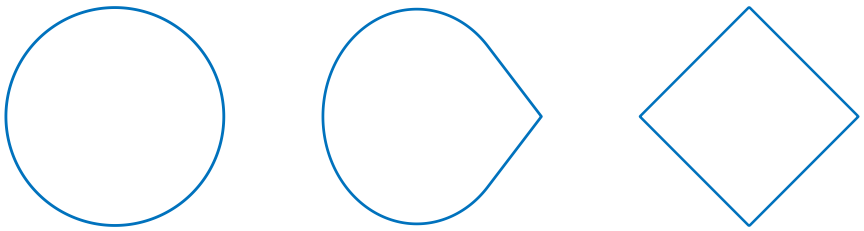


Figure 3.10: Approximating the ℓ_1-ball in $\mathbb{R}^2$ as the projection of the free-spectrahedron in $\mathbb{S}^2$ (left), $\mathbb{S}^3$ (center), and $\mathbb{S}^4$ (right).

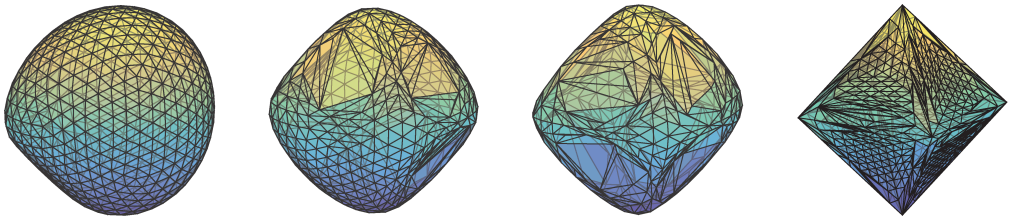


Figure 3.11: Approximating the ℓ_1-ball in $\mathbb{R}^3$ as the projection of free spectrahedron in $\mathbb{S}^3$, $\mathbb{S}^4$, $\mathbb{S}^5$, and $\mathbb{S}^6$ (from left to right).

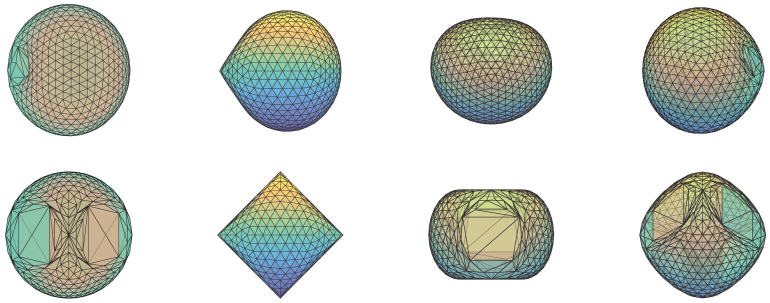


Figure 3.12: Reconstructions of $\mathcal{K}^\star$ (defined in (3.16)) as the projection of O^3 (top row) and O^4 (bottom row). The figures in each row are different views of a single reconstruction, and are orientated in the $(0,0,1)$,$(0,1,0)$,$(1,0,1)$, and $(1,1,0)$ directions (from left to right) respectively.

out some of the 'pointy' features of UPillow; see for example the reconstructions based on O^3 and on O^4 viewed in the $(0,1,0)$ direction in Figure 3.12.

3.5.3 Polyhedral Approximations of the ℓ_2-ball and the Tammes Problem

In the third set of synthetic experiments, we consider reconstructions of the ℓ_2-ball in $\mathbb{R}^3$ via linear images of the simplex; i.e., polytopes. The experimental set-up is similar to the previous series of experiments: we supply 2562 regularly-spaced noiseless support function measurements of the ℓ_2-ball as an input to (3.2), and we select C to be the simplex Δ^q in some larger-dimensional space q, and for q over a range of values.

The purpose of our experiment is to explore a specific instance of the broader question of approximating the ℓ_2-ball in $\mathbb{R}^3$ as a polytope. Such a problem has been widely studied in different contexts and varying forms. For instance, the Tammes problem seeks the optimal placement of q points on $\mathcal{S}^2$ so as to maximize the minimum pairwise distance, and is inspired by pollen patterns [142].[6] A separate body of work studies polyhedral approximations of general compact convex bodies in the asymptotics [25]. Yet another piece of work arising from optimization is that of computing polyhedral approximations of the second-order cone [14] – in particular, the approach in [14] leads to an approximation that is based on expressing the ℓ_2-ball via a nested hierarchy of planar spherical constraints and subsequently approximate these constraints with regular polygons.

Figure 3.13 shows the optimal solutions computed using our method for $q \in \{4, 5, \ldots, 12\}$. We remark that the configurations in our solutions are similar to those of the Tammes Problem [41, 125] in some instances:

$$\operatorname*{argmax}_{\{\mathbf{a}_j\}_{j=1}^q \subset \mathcal{S}^{d-1}} \min_{1 \le k < l \le q} \operatorname{dist}(\mathbf{a}_k, \mathbf{a}_l) = \operatorname*{argmin}_{\{\mathbf{a}_j\}_{j=1}^q \subset \mathcal{S}^{d-1}} \max_{1 \le k < l \le q} \langle \mathbf{a}_k, \mathbf{a}_l \rangle. \tag{3.17}$$

Specifically, the face lattice (as a graph) of our solutions is isomorphic to that of the Tammes for $q \in \{4, 5, 6, 7, 12\}$, which suggests that these configurations are stable and optimal for a broader class of objectives. We are currently not aware if the difference between solutions to both sets of problems for $q \in \{8, 9, 10, 11\}$ arises because our method recovers a solution that is locally optimal but not globally optimal due to a lack of random initializations (in generating these results, we apply 500 initializations for each instance of q), or is inherently due to the different objectives that both problems seek to optimize. We conjecture that the difference for $q = 8$ is due to the latter reason we raised, as an initialization using a configuration that is isomorphic to the Tammes solution led to a suboptimal local minimum.

3.5.4 Reconstruction of a Human Lung

In the final set of experiments we apply our algorithm to reconstruct a convex mesh of a human lung. The purpose of this experiment is to demonstrate the utility of our algorithm in a setting in which the underlying object is not convex. Indeed, in many applications in practice of reconstruction from support function evaluations, the underlying set of interest is not convex; however, due to the nature of the measurements available, one seeks a reconstruction of the convex hull of the

[6] Tammes Problem is a special case of Thompson's Problem, as well as Smale's 7th Problem [133].

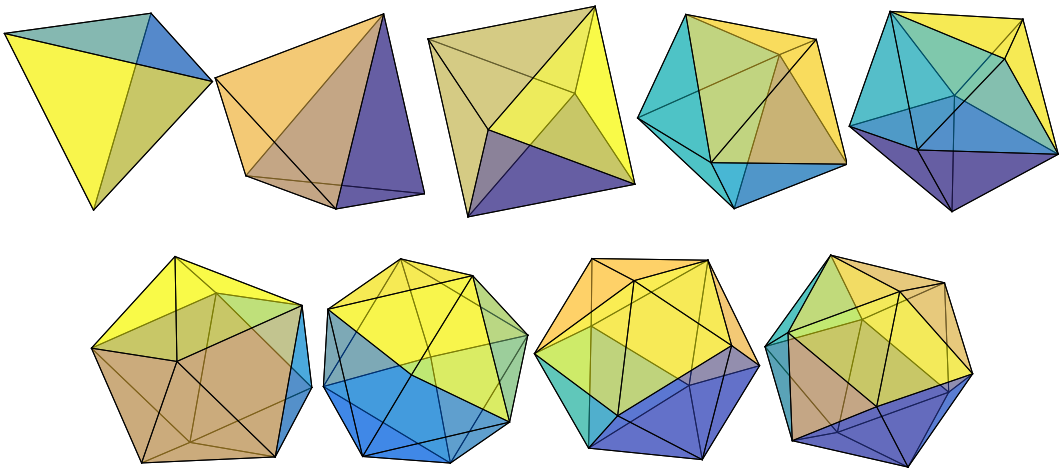
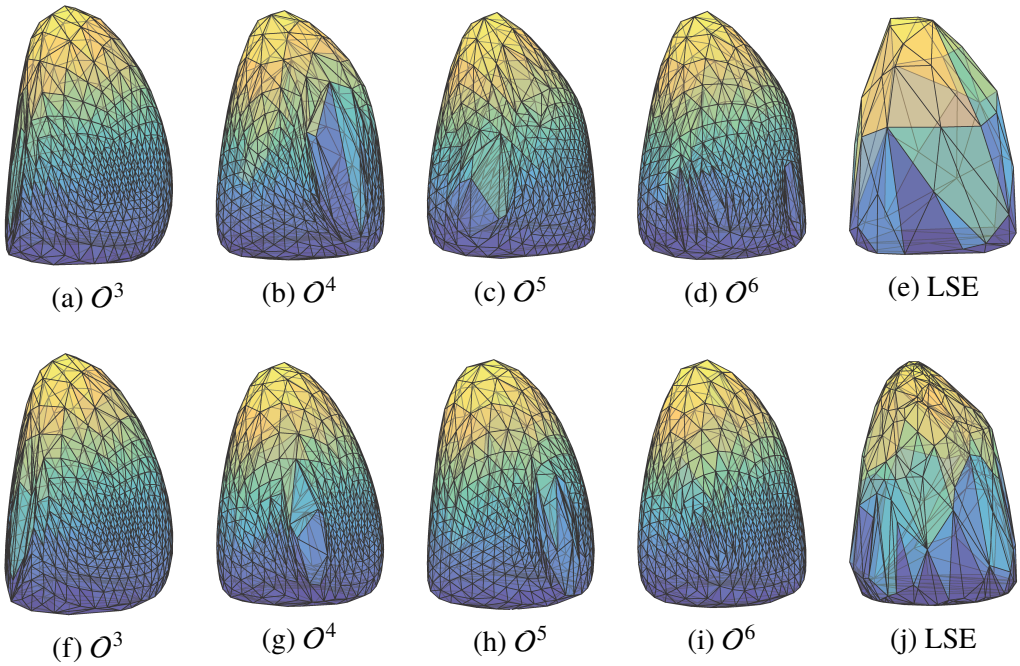


Figure 3.13: Approximating the ℓ_2-ball in $\mathbb{R}^3$ as the projection of Δ^q for $q \in \{4, 5, \ldots, 12\}$ (from left to right, top to bottom).

underlying set. In the present example, the set of interest is obtained from the CT scan of the left lung of a healthy individual [59]. We note that a priori it is unclear whether the convex hull of the lung is well-approximated as the linear image of a low-dimensional simplex or free spectrahedron.

We first obtain $n = 50$ noiseless support function evaluations of the lung (note that this object lies in $\mathbb{R}^3$) in directions that are generated uniformly at random over the sphere $\mathcal{S}^2$. In the top row of Figure 3.14 we show the reconstructions as projections of O^q for $q \in \{3, 4, 5, 6\}$, and we contrast these with the LSE. We repeat the same experiment with $n = 300$ measurements, with the reconstructions shown in the bottom row of Figure 3.14.

To concretely compare the results obtained using our framework and those based on the LSE, we contrast the description complexity – the number of parameters used to specify the reconstruction – of the estimates obtained from both frameworks. An estimator computed using our approach is specified by a projection map $A \in L(\mathbb{R}^q, \mathbb{R}^d)$, and hence it requires dq parameters; while the LSE proposed by the algorithm in [60] assigns a vertex to every measurement, and hence it requires dn parameters. The LSE using $n = 300$ measurements requires 3×300 parameters to specify whereas the estimates obtained using our framework – these are specified as projections of O^5 and O^6 – require 3×15 and 3×21 parameters respectively. Despite requiring significantly fewer parameters to specify, the estimates obtained using our method offer comparable quality to the LSE. This substantial discrepancy highlights the drawback of using polyhedral sets of growing complexity to approximate non-



(a) $\mathcal{O}^3$ (b) $\mathcal{O}^4$ (c) $\mathcal{O}^5$ (d) $\mathcal{O}^6$ (e) LSE

(f) $\mathcal{O}^3$ (g) $\mathcal{O}^4$ (h) $\mathcal{O}^5$ (i) $\mathcal{O}^6$ (j) LSE

Figure 3.14: Reconstructions of the left lung from 50 support function measurements (top row) and 300 support function measurements (bottom row). Subfigures (a),(b),(c),(d),(f),(g),(h), and (i) are projections of free spectrahedra with dimensions as indicated, and subfigures (e) and (j) are LSEs.

polyhedral objects in higher dimensions.

3.6 Conclusions and Future Directions

In this paper we describe a framework for fitting tractable convex sets to noisy support function evaluations. Our approach provides many advantages in comparison to the previous LSE-based methods, most notably in settings in which the measurements available are noisy or small in number as well as those in which the underlying set to be reconstructed is non-polyhedral. We discuss here some potential future directions:

Algorithmic performance guarantees. An important question that merits further investigation is that of designing algorithms that can solve (3.2) to global optimality. The connection between (3.2) and K-means clustering – as discussed in Section 3.1.3.3 – suggests that computing such globally optimal solutions of (3.2) may be computationally intractable as the results in [42, 95] prove that K-means clustering is NP-hard. Accordingly, an approach that is commonly adopted in inverse problems arising in data analysis is to identify subclasses of problem instances for which a family of algorithms succeeds in obtaining globally optimal solutions. In the context

of the algorithms described in Section 3.4, such a strategy may, for example, entail designing suitable initialization methods for the two procedures described there. Alternatively, it is also of interest to derive 'global' methods, such as those based on convex relaxations.

Informed choices of the lifting dimension q. In many settings in practice, a suitable choice of the lifting dimension q is usually not known in advance. Smaller values of q allow us to obtain more concisely described reconstructions, although such estimates may not fit the given data well; on the other hand, larger values of q provide better fidelity to the data and yield more complex reconstructions, but one runs the risk of over-fitting. A practically relevant question in our context is to design methods akin to cross-validation to choose q in a data-driven manner.

We illustrate our ideas with the following stylized experiment. In this instance, we consider reconstructing the ℓ_1-ball in $\mathbb{R}^3$ from 100 measurements corrupted by additive Gaussian noise with standard deviation $\sigma = 0.1$. We partition the measurements into two subsets of equal size. Next, we apply our method with the choice of $C = \Delta^q$ as our lifting set on the first partition, and we evaluate the mean squared error (MSE) of our computed estimator on the second partition. We repeat the process across 50 different random partitions, and over values of q in $\{3, \ldots, 10\}$. The left sub-plot of Figure 3.15 shows the MSE averaged over all partitions. We observe that the error decreases as q increases initially as models that are more expressive allow us to fit to the data better. We observe that the error subsequently remains approximately equal (instead of increasing, as one might expect), and this occurs because our regression restricts to convex sets, which prevents the MSE from growing unboundedly.

We remark that our observations apply more generally. In our second experiment, we consider reconstructing the set $\mathcal{K}_{S3} \subset \mathbb{R}^3$ (defined below) from 200 measurements corrupted by additive Gaussian noise with standard deviation $\sigma = 0.05$. The remaining sequence of steps is identical to the first experiment. The set $\mathcal{K}_{S3}$ is defined as the convex hull of three disjoint planar discs $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3\}$, where each $\mathcal{S}_j$ is defined as follows:

$$\mathcal{S}_j = Q_j \left(\left\{ \begin{pmatrix} \cos\theta \\ 1 \\ \sin\theta \end{pmatrix} : \theta \in \mathbb{R} \right\} \right), \quad Q_j = \begin{pmatrix} \cos(2\pi j/3) & -\sin(2\pi j/3) & \\ \sin(2\pi j/3) & \cos(2\pi j/3) & \\ & & 1 \end{pmatrix}.$$

It is not difficult to see that $\mathcal{K}_{S3}$ is representable as the projection of $\mathcal{O}^6$. Our results

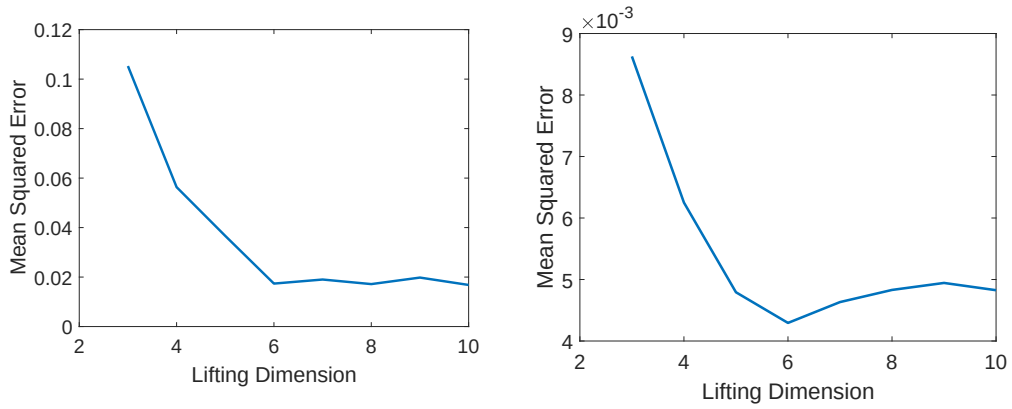


Figure 3.15: Choosing the lifting dimension in a data-driven manner. The left sub-plot shows the cross validation error of reconstructing the ℓ_1-ball in $\mathbb{R}^3$ as the projection of Δ^q over different choices of q, and the right sub-plot shows the same quantity for $\mathcal{K}_{S3} \subset \mathbb{R}^3$ (see accompanying text) as the projection of O^p over different choices of p.

in the right sub-plot of Figure 3.15 indicates a trend that is similar to that of the first experiment.

Richer families of tractable convex sets. As described in the introduction, a restriction in the development of this paper is that, given a lifting dimension q, we do not consider further optimization of the set $C \subset \mathbb{R}^q$. For example, in settings where C is specified as a fixed affine slice of a convex cone (the simplex and the free spectrahedron O^p are specific affine slices of the nonnegative orthant and the cone of positive semidefinite matrices respectively), we do not consider searching over all affine slices of the cone from which C is obtained. The complication that arises with such additional optimization is that the gradient of the support function – with respect to the parameters of the affine space describing the slice – corresponding to such sets is not easily characterized analytically; indeed, requiring such gradient information is in some sense equivalent to asking for a compact description of the sensitivity analysis for arbitrary conic optimization problems, which is not available in general. It would therefore be useful to identify broader yet structured families of sets than the ones we have considered in this paper for which the gradient of the support function continues to be efficiently characterized.

Chapter 4

OPTIMAL APPROXIMATIONS OF CONVEX SETS AS SPECTRAHEDRA

4.1 Introduction

Semidefinite programming is the class of optimization instances in which we minimize linear functionals over feasible regions that are specified as the projection of an affine slice of the cone of positive semidefinite matrices. An outstanding question is to further understand the collection of sets that are representable via semidefinite programming. Progress towards such a goal help us further recognize optimization instances for which we can solve efficiently.

There is a wide spectrum of approaches towards the above goal. One approach is through the lens of convex algebraic geometry, which seeks to understand the algebraic properties of semidefinite-representable sets. A different approach is via the lens of computational complexity, which seeks to understand minimal representations of convex sets as semidefinite programs.

In this chapter, we seek to understand the *expressiveness* of semidefinite representations. More precisely: *Given a compact convex $C^\star \subset \mathbb{R}^d$, what is the optimal approximation (in a suitably defined manner) as a spectrahedron of size p?* Recall that a convex set $C \subset \mathbb{R}^d$ is a *spectrahedron* of size p if it is expressible via a linear matrix inequality involving symmetric matrices of dimensions $p \times p$:

$$C = \{(x_1, \ldots, x_d)' : A_0 + A_1 x_1 + \ldots A_d x_d \succeq 0\}, \quad A_i \in \mathbb{S}^p.$$

In the following, we describe a numerical approach for computing the optimal solution. These are based on applying the ideas in Chapter 3 for reconstructing a convex set from support function evaluations, and by viewing our task of approximating convex sets as spectrahedra in the appropriate dual perspective. Our experiments show that the minimum-sized spectrahedron required to approximate a given set to high precision may sometimes be surprisingly small.

We remark that the nature of our subsequent discussion is exploratory, and it represents work in progress.

4.2 Optimal Approximations of Compact Sets as Spectrahedra of Fixed Size

Let $C^\star \subset \mathbb{R}^d$ be a compact convex set containing the origin in its interior.

Recall that the *gauge function* of a convex set $\gamma_C(\cdot) : \mathbb{R}^d \to [0, \infty]$ is defined as

$$\gamma_C(\mathbf{x}) = \inf\{t : \mathbf{x} \in t \cdot C\}.$$

The *radial function* of a convex set C is the reciprocal of its gauge function, and is defined for all non-zero vectors $\mathbf{x} \in \mathbb{R}^d$ [124]. There is a very natural geometric interpretation of the radial function; namely, it gives the scale of the vector $\mathbf{x}$ such that it remains in C

$$1/\gamma_C(\mathbf{x}) = \sup\{t : t\,\mathbf{x} \in C\}.$$

In particular, if $\mathbf{u} \in \mathbb{R}^d$ is also unit Euclidean-norm, then the radial function measures the distance between the origin and the boundary of C in the positive $\mathbf{u}$ direction.

In this chapter we consider computing a spectrahedron of size p that minimizes the squared error loss in terms of its radial function

$$\arg\min_{C} \int \left(\frac{1}{\gamma_{C^\star}(\mathbf{u})} - \frac{1}{\gamma_C(\mathbf{u})}\right)^2 \mu(d\mathbf{u}), \quad C \text{ is a spectrahedron of size } p. \tag{4.1}$$

Here, $\mu(\cdot)$ denotes the uniform surface measure over the unit sphere $\mathcal{S}^{d-1}$.

Remark. *Notice that we suffer no loss of generality in assuming that $C^\star$ contains the origin in its interior – there is clearly no loss in assuming that $C^\star$ is full-dimensional, and if $C^\star$ does not contain the origin, we simply apply a translation.*

The dual perspective. We begin by describing a reformulation of the optimization instance (4.1), which is based on parameterizing the variable C in its *polar*.

Formally, given a set $C \subset \mathbb{R}^d$, the polar set $C^\circ \subset \mathbb{R}^d$ is defined as

$$C^\circ := \{\mathbf{y} : \langle \mathbf{x}, \mathbf{y} \rangle \leq 1 \text{ for all } \mathbf{x} \in C\} \subset \mathbb{R}^d.$$

Polarity defines a duality over the collection of compact convex sets that contains the origin in its interior. In particular, polarity transforms information about the boundary in the ambient space into support information in the dual space. More precisely, suppose that $C \subset \mathbb{R}^d$ is compact convex, and it satisfies $\mathbf{0} \in \text{int}(C)$. Then $C^{\circ\circ} = C$, and $\gamma_C(\cdot) = h_{C^\circ}(\cdot)$ (see, for instance, Theorem 1.6.7 in [124]). Here, $h_C(\mathbf{x}) := \sup_{\mathbf{g} \in C} \langle \mathbf{g}, \mathbf{x} \rangle$ denotes the support function of C evaluated at $\mathbf{x}$.

Proposition 4.2.1 *Define*

$$\hat{C}_P := \arg\min_{C} \int \left(\frac{1}{\gamma_{C^\star}(\mathbf{u})} - \frac{1}{\gamma_C(\mathbf{u})} \right)^2 \mu(d\mathbf{u}), \quad C \text{ is a spectrahedron of size } p, \tag{4.2}$$

and define

$$\hat{C}_D := \arg\min_{\mathcal{L} \in L(\mathbb{S}^p, \mathbb{R}^d)} \int \left(\frac{1}{\gamma_{C^\star}(\mathbf{u})} - \frac{1}{h_{\mathcal{L}(O^p)}(\mathbf{u})} \right)^2 \mu(d\mathbf{u}). \tag{4.3}$$

Suppose that $\hat{C}_P$ and $\hat{C}_D$ are both compact, and they contain the origin in their respective interiors. In addition, suppose that the solutions $\hat{C}_P$ and $\hat{C}_D$ are respectively unique. Then $\hat{C}_P = \hat{C}_D$.

Here, $O^p = \mathbb{S}^p_+ \cap \{X : \langle I, X \rangle = 1\}$ denotes the free spectrahedron with dimensions $p \times p$.

The proof of Proposition 4.2.1 relies on some basic properties about the polarity operation. We begin by noting that every spectrahedron of size p containing the origin in its interior can be expressed as the following

$$\{\mathbf{x} : I \succeq \mathcal{L}'\mathbf{x}\} \subset \mathbb{R}^d$$

for some linear map $\mathcal{L} \in L(\mathbb{S}^p, \mathbb{R}^d)$.

Next we compute the polar of the above set.

Proposition 4.2.2 *Let $C = \{\mathbf{x} : I \succeq \mathcal{L}'\mathbf{x}\}$ for some $\mathcal{L} \in L(\mathbb{S}^p, \mathbb{R}^d)$. Then the polar set C° is given by*

$$C^\circ = \{\mathcal{L}(X) : X \in \mathbb{S}^p_+, \langle I, X \rangle \leq 1\} = \mathcal{L}(\mathbb{S}^p_+ \cap \{X : \langle I, X \rangle \leq 1\}).$$

In particular, if $\mathbf{0} \in \mathcal{L}(\mathbb{S}^p_+ \cap \{X : \langle I, X \rangle \leq 1\})$, then $C^\circ = \mathcal{L}(O^p)$.

Proof of Proposition 4.2.2. It follows from the definition of polarity that

$$C^\circ := \{\mathbf{y} : \langle \mathbf{y}, \mathbf{x} \rangle \leq 1 \text{ for all } \mathbf{x} \in C\}$$

This can be equivalently expressed in terms of the solution of the following convex program

$$C^\circ = \left\{ \mathbf{y} : \sup_{\mathbf{x}: I \succeq \mathcal{L}'\mathbf{x}} \langle \mathbf{x}, \mathbf{y} \rangle \leq 1 \right\}.$$

It is clear that the origin is an interior point of the feasible region. Hence Slater's condition is satisfied. By strong duality, we have

$$\begin{aligned} C^\circ &= \left\{\mathbf{y} : \exists\, X, \langle I, X\rangle \le 1, \mathcal{L}(X) = \mathbf{y}, X \in \mathbb{S}^p_+\right\} \\ &= \{\mathcal{L}(X) : X \in \mathbb{S}^p_+, \langle I, X\rangle \le 1\} \\ &= \mathcal{L}(\mathbb{S}^p_+ \cap \{X : \langle I, X\rangle \le 1\}). \end{aligned}$$

In fact, one can further write

$$C^\circ = \bigcup_{0 \le t \le 1} t \cdot \mathcal{L}(\mathcal{O}^p).$$

It is clear that if $\mathbf{0} \in \mathcal{L}(\mathcal{O}^p)$ then $C^\circ = \mathcal{L}(\mathcal{O}^p)$. □

Proof of Proposition 4.2.1. Since $\hat{C}_P$ contains the origin in its interior, the optimal solution may be expressed as

$$\hat{C}_P = \{\mathbf{x} : I \succeq \hat{\mathcal{L}}'_P \mathbf{x}\},$$

for some linear map $\hat{\mathcal{L}}_P \in L(\mathbb{S}^p, \mathbb{R}^d)$. By Proposition 4.2.2, the polar set $\hat{C}^\circ_P$ is given by

$$\hat{C}^\circ_P = \hat{\mathcal{L}}_P(\mathcal{O}^p).$$

Next, let $\hat{C}_D = \hat{\mathcal{L}}_D(\mathcal{O}^p)$. It is clearly closed. Since $\mathbf{0} \in \mathrm{int}(\hat{C}_D)$, it follows that $\hat{C}^\circ_D$ is compact. Furthermore, by Proposition 4.2.2, we have $\hat{C}^\circ_D = \{\mathbf{x} : I \succeq \hat{\mathcal{L}}'_D \mathbf{x}\}$.

We may conclude that $\hat{C}_D = \hat{C}_P$. Suppose that this is not the case. Then it must be that $\hat{\mathcal{L}}_P$ defines a suboptimal solution in (4.3). However, the set $\hat{C}^\circ_D$ would attain then a value in (4.2) that is strictly lower that the unique optimum $\hat{C}_P$, which is not possible. So they must both attain the same errors. By uniqueness, we further conclude that $\hat{C}_D = \hat{C}_P$. □

4.3 Algorithms

In this section we describe our algorithms for minimizing (4.3). Our first algorithm is based on gradient descent and our second algorithm is a form of alternating minimization.

First we note that the minimization instance (4.3) requires the computation of a surface integral, which is not feasible to compute in general. Our approach is to approximate the integral as point masses on the surface, which leads to the following

optimization instance

$$\underset{\mathcal{L}\in L(\mathbb{R}^q,\mathbb{R}^d)}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)}-\frac{1}{h_{\mathcal{L}(\mathcal{D})}(\mathbf{u}^{(i)})}\right)^2. \tag{4.4}$$

Gradient Descent. Our first approach is based on gradient descent. The derivative of objective in (4.4) with respect to the linear map $\mathcal{L}$ (provided it exists) is given by

$$\frac{2}{n}\sum_{i=1}^{n}\left(y^{(i)}-1/h_{h_{\mathcal{D}}}(\mathcal{L}'\mathbf{u}^{(i)})\right)\left(\frac{1}{h_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)})}\right)^2\mathbf{u}^{(i)}\otimes E_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)}).$$

Here, $E_{\mathcal{D}}(X)$ is the unit-norm positive semidefinite matrix that corresponds to the largest eigenvalue of X. We summarize the steps in Algorithm 6.

Algorithm 6 Convex Set Regression via Gradient Descent

Input: A collection $\{(\mathbf{u}^{(i)},y^{(i)})\}_{i=1}^n\subset\mathbb{R}^d\times\mathbb{R}$ of support function evaluations; a compact convex set $\mathcal{D}\subset\mathbb{R}^q$; an initialization $\mathcal{L}\in L(\mathbb{R}^q,\mathbb{R}^d)$; a step size $\eta\in(0,1]$
Algorithm: Repeat until convergence
1. $D\leftarrow\frac{2}{n}\sum_{i=1}^n(y^{(i)}-1/h_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)}))(1/h_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)}))^2\mathbf{u}^{(i)}\otimes E_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)})$
2. $\mathcal{L}\leftarrow\mathcal{L}-\eta D$
Output: Final iterate $\mathcal{L}$

Although the algorithm presented is based on a fixed stepsize, in practice stepsizes may also be chosen via some type of line search.

Alternating Minimization. The second algorithm is based on iteratively computing solutions to linearizations of the first-order optimality conditions of (4.4). By noting that $h_{\mathcal{D}}(\mathcal{L}'\mathbf{u})=\langle\mathcal{L}'\mathbf{u},E_{\mathcal{D}}(\mathcal{L}'\mathbf{u})\rangle=\langle\mathcal{L},\mathbf{u}\otimes E_{\mathcal{D}}(\mathcal{L}'\mathbf{u})\rangle$, the first-order optimality conditions of (4.4) may be expressed as follows:

$$\begin{aligned}0 &= \frac{2}{n}\sum_{i=1}^{n}\left(y^{(i)}h_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)})-1\right)\left(\frac{1}{h_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)})}\right)^3\mathbf{u}^{(i)}\otimes E_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)})\\ &= \frac{2}{n}\sum_{i=1}^{n}\left(y^{(i)}\langle\mathcal{L},\mathbf{u}^{(i)}\otimes E_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)})\rangle-1\right)\left(\frac{1}{h_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)})}\right)^3\mathbf{u}^{(i)}\otimes E_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)}).\end{aligned}$$

Our algorithm suffers similar conditioning issues as those in Section 3.4 of Chapter 3. To address these issues, we apply an intermediate Tikhonov regularization analogous to that of Algorithm 5.

4.4 Numerical Experiments

TV-screen. Our first example is the set $C_1:=\{(x,y):x^4+y^4\leq 1\}\subset\mathbb{R}^2$, which is also known as the *TV-screen* (see Figure 4.1). The set is C_1 is *not* expressible as

Algorithm 7 Convex Regression via Alternating Minimization

Input: A collection $\{(\mathbf{u}^{(i)}, y^{(i)})\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ of support function evaluations; a compact convex set $\mathcal{D} \subset \mathbb{R}^q$; an initialization $\mathcal{L} \in L(\mathbb{R}^q, \mathbb{R}^d)$; a choice of regularization parameter $\gamma > 0$

Algorithm: Repeat until convergence

1. $\boldsymbol{e}^{(i)} \leftarrow E_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)})$, $h^{(i)} \leftarrow h_{\mathcal{D}}(\mathcal{L}'\mathbf{u}^{(i)})$

2. $V \leftarrow \frac{1}{n}\sum_{i=1}^n y^{(i)}(h^{(i)})^{-3}((\mathbf{u}^{(i)} \otimes \boldsymbol{e}^{(i)}) \otimes (\mathbf{u}^{(i)} \otimes \boldsymbol{e}^{(i)}))$, $\quad Y \leftarrow \frac{1}{n}\sum_{i=1}^n (h^{(i)})^{-3}(\mathbf{u}^{(i)} \otimes \boldsymbol{e}^{(i)})$

3. $\mathcal{L} \leftarrow (V + \gamma I)^{-1}(Y + \gamma \mathcal{L})$

Output: Final iterate $\mathcal{L}$

a spectrahedron [77], though it is expressible as the projection of an affine slice of the PSD cone.

In Figure 4.2 we show the approximations of C_1 computed using Algorithm 7 as spectrahedra of sizes 2, 3, and 4. In this instance, a spectrahedron of size 4 is sufficient to approximate C_1 up to a mean squared error of 1.7×10^{-5}.

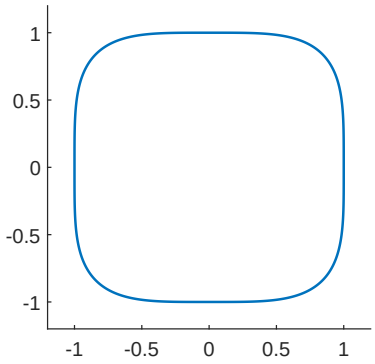
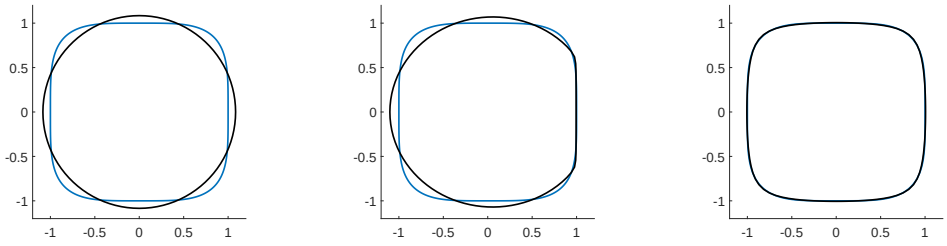


Figure 4.1: $\{(x, y) : x^4 + y^4 \leq 1\}$, also known as the TV-screen.



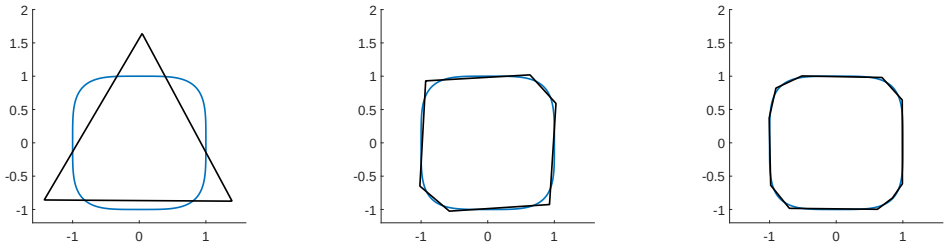
(a) Size 2 (b) Size 3 (c) Size 4

Figure 4.2: Approximations of the TV-screen as spectrahedra.

We compare our reconstructions with polyhedral approximation of comparable complexity. In Figure 4.3 we show the approximations of C_1 as polyhedra with 3, 6,

and 10 facets. The number of facets are chosen to match the number of parameters required to specify spectrahedra of sizes 2, 3, and 4.

We compare the mean squared errors of both families of reconstructions in Figure 4.1. Notice that although the spectrahedral approximation of size 4 requires as many parameters to specify as the corresponding polyhedral approximation with 10 facets, the latter attains a fit with MSE that is approximately a ninth of former.



(a) 3 facets (b) 6 facets (c) 10 facets

Figure 4.3: Approximations of the TV-screen as polyhedra.

Degrees of freedom	2×3	2×6	2×10
Polyhedron	5.5×10^{-2}	9.9×10^{-4}	1.5×10^{-4}
Spectrahedron	4.4×10^{-3}	3.1×10^{-3}	1.7×10^{-5}

Table 4.1: Mean Squared Errors of approximations of the TV-screen as polyhedra and spectrahedra of different sizes.

A non-semialgebraic set in $\mathbb{R}^2$. Our second example is a set in $\mathbb{R}^2$ bounded by three curves

$$C_2 := \{(x, y) : x \leq 1/2, y \leq 1/2, \exp(-3/2 - x) - 1/2 \leq y\}.$$

We show the set C_2 in Figure 4.4. The set C_2 is *not* semialgebraic because a part of its boundary is defined by the exponential function. Consequently, it is not expressible via semidefinite programming.

In Figure 4.5 we show the approximations of C_2 as spectrahedra of sizes 2, 3, and 4. In this instance, a spectrahedron of size 4 is sufficient to obtain an approximation with a MSE of size $\approx 1.0 \times 10^{-8}$ (see Figure 4.6).

A semialgebraic set in $\mathbb{R}^3$. Our third example considers the convex hull of the following variety

$$\mathcal{V} := \{(x, y, z) : x^4 - y^2 - z^2 = 0, x^4 + y^2 + z^2 - 1 = 0\}.$$

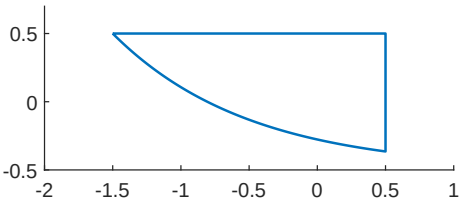
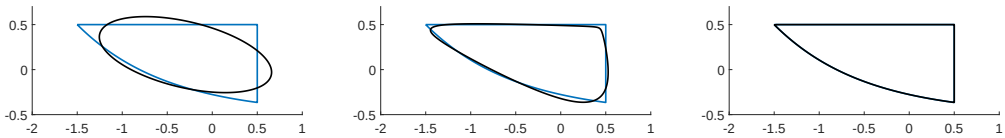


Figure 4.4: An non-semialgebraic set defined as $\{(x, y) : x \leq 1/2, y \leq 1/2, \exp(-3/2 - x) - 1/2 \leq y\}$.



(a) Size 2 (b) Size 3 (c) Size 4

Figure 4.5: Approximations of C_2 as spectrahedra.

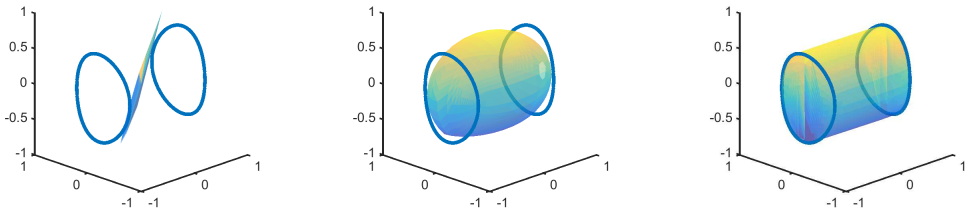
Size of spectrahedra	2	3	4
Mean Squared Error	1.3×10^{-2}	9.5×10^{-4}	9.6×10^{-9}

Figure 4.6: Mean Squared Errors of approximations of C_2 as spectrahedra of different sizes.

This set is based on a worked example in [68].

In Figure 4.7 we show the approximations of conv($\mathcal{V}$) as spectrahedra of sizes 2, 3, and 4 computed using our method. We observe that a spectrahedra of size 4 is sufficient to obtain a MSE $\lesssim 10^{-6}$ (see Figure 4.8).

To emphasize the compactness of our description, we compare our approximations with the outer approximations obtained using *theta bodies relaxations* [69]. The latter approach offers a principled manner for constructing convex outer approximations of semialgebraic sets that are representable via semidefinite programming. These relaxations are presented as a hierarchy that converge towards the convex hull. Based on the computations in [68], the first theta body relaxation is an ellipsoid, and the second theta body is a convex set that can be expressed as a semidefinite



(a) Size 2 (b) Size 3 (c) Size 4

Figure 4.7: Reconstructions of conv($\mathcal{V}$) using our method. The variety $\mathcal{V}$ is outlined in blue.

Size of spectrahedra	2	3	4
Mean Squared Error	6.1×10^{-1}	5.7×10^{-3}	2.1×10^{-7}

Figure 4.8: Mean Squared Errors of approximations of conv($\mathcal{V}$) as spectrahedra

program in dimensions 9. The numerical results in [68] do further suggest that the second theta body relaxation is *exact*. Although our reconstruction as spectrahedron of size 4 is approximate, it is able to achieve a high degree of accuracy using a far smaller semidefinite program.

Chapter 5

HIGH-DIMENSIONAL CHANGE-POINT ESTIMATION: COMBINING FILTERING WITH CONVEX OPTIMIZATION

5.1 Introduction

Change-point estimation is the identification of abrupt changes or anomalies in a sequence of observations. Such problems arise in numerous applications such as product quality control, data segmentation, network analysis, and financial modeling; an overview of the change-point estimation literature can be found in [13, 39, 114, 149]. As in other inferential tasks encountered in contemporary settings, a key challenge underlying many modern change-point estimation problems is the increasingly large dimensionality of the underlying sequence of signals – that is, the signal at each location in the sequence is not scalar-valued but rather lies in a high-dimensional space. This challenge leads both to computational difficulties as well as to complications with obtaining statistical consistency in settings in which one has access to a small number of observations (relative to the dimensionality of the space in which these observations live).

A prominent family of methods for estimating the locations of change-points in a sequence of noisy scalar-valued observations is based on the *filtered derivative* approach [7, 13, 15, 17, 18]. Broadly speaking, these procedures begin with an application of a low-pass filter to the sequence, followed by a computation of pairwise differences between successive elements, and finally the implementation of a thresholding step to estimate change-points. A large body of prior literature has analyzed the performance of this family of algorithms and its variants [7, 17, 18]. Unfortunately, as we describe in Section 5.3, the natural extension of this procedure to the high-dimensional setting leads to performance guarantees for reliable change-point estimation that require the underlying signal to remain unchanged for long portions of the sequence. Such requirements tend to be unrealistic in applications such as financial modeling and network analysis in which rapid transitions in the underlying phenomena trigger frequent changes in the associated signal sequences.

5.1.1 Our contributions

To alleviate these difficulties, modern signal processing methods for high-dimensional data – in a range of statistical inference tasks such as denoising [19, 47, 63], model

selection [29, 99, 150], the estimation of large covariance matrices [21, 22], and others [32, 35, 48, 56, 116] – recognize and exploit the observation that signals lying in high-dimensional spaces typically possess low-dimensional structure. For example, images frequently admit sparse representations in an appropriately transformed domain [26, 32] (e.g., the wavelet domain), while covariance matrices are well-approximated as low-rank matrices in many settings (e.g., correlations between financial assets). The exploitation of low-dimensional structure in solving problems such as denoising leads to consistency guarantees that depend on the intrinsic low-dimensional "complexity" of the data rather than on the ambient (large) dimension of the space in which they live. A notable feature of several of these structure-exploiting procedures is that they are based on convex optimization methods, which can lead to tractable numerical algorithms for large-scale problems as well as to insightful statistical performance analyses. Motivated by these ideas, we propose a new approach for change-point estimation in high dimensions by integrating a convex optimization step into the filtered derivative framework (see Section 5.3). We prove that the resulting method provides reliable change-point estimation performance in high-dimensional settings, with guarantees that depend on the underlying low-dimensional structure in the sequence of observations rather than on their ambient dimension.

To illustrate our ideas and arguments concretely, we consider a setup in which we are given a sequence $\mathbf{y}[t] \in \mathbb{R}^q$ for $t = 1, \ldots, n$ of observations of the form:

$$\mathbf{y}[t] = \mathbf{x}^\star[t] + \boldsymbol{\varepsilon}[t]. \tag{5.1}$$

Here $\mathbf{x}^\star[t] \in \mathbb{R}^q$ is the underlying signal and the noise is independent and identically distributed across time as $\boldsymbol{\varepsilon}[t] \sim \mathcal{N}(0, \sigma^2 I_{q\times q})$. The signal sequence $\mathscr{X} := \{\mathbf{x}^\star[t]\}_{t=1}^n$ is assumed to be piecewise constant with respect to t. The set of change-points is denoted by $\tau^\star \subset \{1, \ldots, n\}$, i.e., $t \in \tau^\star \Leftrightarrow \mathbf{x}^\star[t] \neq \mathbf{x}^\star[t+1]$, and the objective is to estimate the set $\tau^\star$. A central aspect of our setup is that each $\mathbf{x}^\star[t]$ is modeled as having an efficient representation as a linear combination of a small number of elements from a known set $\mathscr{A}$ of building blocks or atoms [19, 30, 32, 33, 35, 47, 48, 53, 116, 143]. This notion of signal structure includes widely studied models in which signals are specified by sparse vectors and low-rank matrices. It also encompasses several others such as low-rank tensors, orthogonal matrices, and permutation matrices. The convex optimization step in our approach exploits knowledge of the atomic set $\mathscr{A}$; specifically, the algorithm described in Section 5.3.2 consists of a denoising operation in which the underlying signal is

estimated from local averages of the sequence $\mathbf{y}[t]$ using a proximal operator based on the atomic norm associated to $\mathscr{A}$ [19, 35, 47]. The main technical result of this paper is that the method we propose in Section 5.3 provides accurate estimates of the change-points $\tau^\star$ with high probability under the condition:

$$\Delta_{\min}^2 T_{\min} \gtrsim \sigma^2 \{\eta^2(\mathscr{X}) + \log n\}, \tag{5.2}$$

where $\Delta_{\min}$ denotes the size (in ℓ_2-norm) of the smallest change among all change-points, $T_{\min}$ denotes the smallest interval between successive change-points, and n is the number of observations. The quantity $\eta(\mathscr{X})$ captures the low-dimensional complexity in the signal sequence $\mathscr{X} := \{\mathbf{x}^\star[t]\}_{t=1}^n$ via a Gaussian distance/width characterization, and it appears in our result due to the incorporation of the convex optimization step. In the high-dimensional setting, the parameter η^2 plays a crucial role as it reflects the underlying low-dimensional structure in the signal sequence of interest; as such it is usually much smaller than the ambient dimension q (we quantify the comparisons in Section 5.2). Indeed, directly applying the filtered derivative method without incorporating a convex optimization step that exploits the signal structure would lead to weaker performance guarantees, with the quantity η^2 in the performance guarantee (5.2) being replaced by the ambient dimension q.

The performance guarantee (5.2) highlights a number of tradeoffs in high-dimensional change-point estimation that result from using our approach. For example, the appearance of the term $\Delta_{\min}^2 T_{\min}$ on the left hand side of (5.2) implies that it is possible to compensate for one of these quantities being small if the other one is suitably large. Further, our algorithm also operates in a causal manner on small portions of the sequence at any given time rather than on the entire sequence simultaneously, and it is therefore useful in 'online' settings. This feature of our method combined with the result (5.2) leads to a more subtle tradeoff between the computational efficiency of the approach and the number of observations n; specifically, our algorithm can be adapted to process larger datasets (e.g., settings in which observations are obtained via high-frequency sampling, leading to larger n) more efficiently without loss in statistical performance by employing a suitable form of convex relaxation based on the ideas discussed in [33]. We discuss these points in greater detail in Section 5.4.

5.1.2 Related work

A recent paper by Harchaoui and Lévy-Leduc [76] is closest in spirit to ours; they describe a convex programming method based on total-variation minimization to detect changes in sequences of scalar-valued signals, and they provide a change-point

estimation guarantee of the form (5.2). Specifically, by combining assumptions (A2) and (A3) in Proposition 3 of [76], the authors show that their algorithm provides accurate estimates of the change-points in the regime $\Delta_{\min}^2 T_{\min}/\log n \to \infty$ as $n \to \infty$, which is similar to our result (5.2) when specialized to scalar-valued signals. In addition to the restriction to scalar-valued signals, the technique in [76] requires knowledge of the full sequence of observations in advance. As a result it is not directly applicable in high-dimensional and online settings unlike our proposed approach.

High-dimensional change-point estimation has received much attention in recent years based on different types of extensions of the scalar case. The diversity of these generalizations of the scalar setting is due to the wide range of applications in which change-point estimation problems arise, each with a unique set of considerations. For example, several papers [38, 52] investigate high-dimensional change-point estimation in settings in which the changes only occur in a small subset of components. Therefore, assumptions about low-dimensional structure are made with regards to the pattern of changes rather than in the signal itself at each time instance (as in our setup). Xie et al. [151] consider a high-dimensional change-point estimation problem in which the underlying signals are modeled as lying on a low-dimensional manifold; although this setup is similar to ours, their algorithmic approach is based on projections onto manifolds rather than on convex optimization, and the types of guarantees obtained in [151] are qualitatively quite different in comparison to (5.2). We also note recent work by Aston and Kirch on high-dimensional change-point problems in which they study the impact of projections on the performance of classical algorithms such as the cumulative-sum method [10]. In the setting where the direction of change is known, the authors demonstrate that a projection along the direction of change yields an algorithm with a recovery guarantee that is independent of the ambient dimension, and is robust to misspecification of the noise covariance.

5.1.3 Paper outline

Section 5.2 gives the relevant background on structured signals that are concisely represented with respect to sets of elementary atoms as well as the analytical tools that are used in the remainder of the paper. In Section 5.3 we describe our algorithm for high-dimensional change-point estimation, and we state the main recovery guarantee of the procedure. In Section 5.4 we discuss the tradeoffs that result from using our approach, and their utility in adapting our algorithm to address challenges beyond high-dimensionality that arise in applications involving change-point esti-

mation. We verify our theoretical results with numerical experiments on synthetic data in Section 5.5, and we conclude with brief remarks and further directions in Section 5.6. The proofs are given in the Appendix.

5.2 Background on Structured Signal Models

5.2.1 Efficient representations with respect to atomic sets

We outline a framework with roots in nonlinear approximation [12, 45, 83, 111] that generalizes several types of low-dimensional models considered in the literature such as sparse vectors and low-rank matrices [19, 21, 22, 32, 35, 48, 99, 106, 116].

Let $\mathscr{A} \subseteq \mathbb{R}^q$ be a compact set that specifies a collection of atoms. We say that a signal $\mathbf{x} \in \mathbb{R}^q$ has a concise representation with respect to $\mathscr{A}$ if it admits a decomposition as a sum of a small number of atoms in $\mathscr{A}$, that is, we are able to write

$$\mathbf{x} = \sum_{i=1}^{s} c_i \mathbf{a}_i, \mathbf{a}_i \in \mathscr{A}, c_i \geq 0, \tag{5.3}$$

for some $s \ll q$. Sparse vectors and low-rank matrices are examples of low-dimensional representations that are expressible in this framework. Specifically, an atomic set for sparse vectors is the set of signed standard basis vectors $\mathscr{A} = \{\pm \mathbf{e}_i | 1 \leq i \leq q\}$, while a natural atomic set for low-rank matrices is set of all rank-one matrices with unit Euclidean norm. Other examples include binary vectors (e.g., in knapsack problems [97]), permutation matrices (in ranking problems [81]), low-rank tensors [87], and orthogonal matrices. Such classes of signals that have concise representations with respect to general atomic sets were studied in the context of linear inverse problems [35], and subsequently in the setting of statistical denoising [19, 33].

In comparison with alternative notions of low-dimensional structure, e.g., manifold models [151], which have been considered previously in the context of high-dimensional change-point estimation (and more generally in signal processing), the setup described here has the virtue that one can employ efficient algorithms for convex optimization methods, and one can appeal to insights from convex geometry in developing and analyzing algorithms for high-dimensional change-point estimation. We discuss the relevant concepts in the next two subsections.

5.2.2 Minkowski functional and proximal operators

A key feature of our change-point estimation algorithm is the incorporation of a signal denoising step that exploits knowledge of the atomic set $\mathscr{A}$. To formally define

the denoising operation, we consider the Minkowski functional $\|\cdot\|_C : \mathbb{R}^q \to [0, \infty]$

$$\|\mathbf{x}\|_C := \inf\{t : \mathbf{x} \in tC,\ t > 0\}, \tag{5.4}$$

defined with respect to a convex set $C \subset \mathbb{R}^q$ such that $\mathscr{A} \subseteq C$, as discussed in [35]. As C is convex, the Minkowski functional $\|\cdot\|_C$ is also convex. This function is also called the gauge function in the convex analysis literature [120]. For a given $\mathbf{y} \in \mathbb{R}^q$ and a convex set C, we consider denoisers specified in terms of the following *proximal operator*:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^q} \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_{\ell_2}^2 + \lambda\|\mathbf{x}\|_C. \tag{5.5}$$

As $\|\cdot\|_C$ is a convex function, this optimization problem is a convex program. To obtain a proximal operator that enforces signal structure in the denoising operation, the set C is usually taken to be the tightest convex set containing the atomic set $\mathscr{A}$, i.e., $C = \text{conv}(\mathscr{A})$. With $C = \text{conv}(\mathscr{A})$, the resulting Minkowski functional is called the *atomic norm*[1] with respect to $\mathscr{A}$, and the associated proximal operator (5.5) is called *atomic norm thresholding* [19]. The atomic norm has been studied in the approximation theory literature for characterizing approximation rates associated with best k-term approximants [12, 45, 83, 111], and its convex-geometric properties were investigated in [35] in the context of ill-posed linear inverse problems. When $\mathscr{A} = \{\pm\mathbf{e}_i | 1 \le i \le q\}$ is the collection of signed standard basis vectors, the atomic norm with respect to $\mathscr{A}$ is simply the ℓ_1-norm in $\mathbb{R}^q$. Similarly, the atomic norm corresponding to unit-Euclidean-norm rank-one matrices is the matrix nuclear norm. More generally, one can define atomic norms associated to other types of structured objects such as permutation matrices, low-rank tensors, orthogonal matrices, and signed vectors; see [35] for a detailed list. Atomic norm thresholding naturally generalizes soft-thresholding based on the ℓ_1-norm for sparse signals to a more general denoising operation for the types of structured signals described here.

One exception to the rule of thumb of choosing $C = \text{conv}(\mathscr{A})$ arises if the atomic norm is intractable to represent, e.g., the tensor nuclear norm [79]. That is, although these norms are convex functions, computing them may in general be computationally intractable. To overcome such difficulties, a natural approach described in [33, 35] is to consider Minkowski functionals of convex sets C that contain $\mathscr{A}$ and that are efficient to represent, i.e., further tractable convex relaxations of $\text{conv}(\mathscr{A})$.

[1] For (5.4) to formally define a norm, we would also need the set $\mathscr{A}$ to be centrally symmetric. Nevertheless, the results in the remainder of the paper hold without this condition, so we use "norm" with an abuse of terminology.

Finally, to avoid dealing with technicalities in degenerate cases, we assume throughout the remainder of the paper that the set $\mathrm{conv}(\mathscr{A}) \subset \mathbb{R}^q$ is a solid convex set containing the origin in its interior. Consequently, we have that $\|\mathbf{x}\|_C < \infty$ for all $\mathbf{x} \in \mathbb{R}^q$.

5.2.3 Summary parameters in signal denoising

Next we describe the relevant convex-geometric concepts for analyzing the performance of proximal denoising operators. For $\mathbf{x} \in \mathbb{R}^q$, the *Gaussian distance* $\eta_C(\mathbf{x})$ [56, 106] with respect to a norm $\|\cdot\|_C$ is defined as

$$\eta_C(\mathbf{x}) := \inf_{\lambda \geq 0} \left\{ \mathop{\mathbb{E}}_{\mathbf{g} \sim \mathcal{N}(0, I_{q \times q})} [\mathrm{dist}(\mathbf{g}, \lambda \cdot \partial \|\mathbf{x}\|_C)] \right\}. \tag{5.6}$$

Here $\mathrm{dist}(\mathbf{g}, \partial\|\mathbf{x}\|_C) := \inf_{\mathbf{w} \in \partial\|\mathbf{x}\|_C} \|\mathbf{w} - \mathbf{g}\|_{\ell_2}$ denotes the distance of $\mathbf{g}$ from the set $\partial\|\mathbf{x}\|_C$, where $\partial\|\mathbf{x}\|_C$ is the subdifferential of the function $\|\cdot\|_C$ at the point $\mathbf{x}$ [120]. We relate the Gaussian distance to the Gaussian width [66] in Appendix C.1 by extending a result in [56].

The Gaussian distance $\eta_C(\mathbf{x})$ is useful for characterizing the performance of the proximal denoising operator (5.5) [19, 33, 106]. Specifically, suppose $\hat{\mathbf{x}} = \arg\min_{\mathbf{x} \in \mathbb{R}^q} \frac{1}{2}\|\mathbf{x}^\star + \boldsymbol{\varepsilon} - \mathbf{x}\|_{\ell_2}^2 + \lambda\|\mathbf{x}\|_C$, then the error between $\hat{\mathbf{x}}$ and $\mathbf{x}^\star$ is bounded as [106]:

$$\|\mathbf{x}^\star - \hat{\mathbf{x}}\|_{\ell_2} \leq \mathrm{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial\|\mathbf{x}^\star\|_C).$$

Taking expectations with respect to $\boldsymbol{\varepsilon}$ and subsequently optimizing the resulting bound with respect to λ yields the Gaussian distance (5.6). We prove a generalization of this result in Appendix C.2, which is relevant to the analysis of the change-point estimation algorithm proposed in Section 5.3.2.

As we will discuss in Section 5.3, the combination of the proximal denoising operator with a suitable filtering step leads to a change-point estimation procedure with performance guarantees in terms of $\eta_C(\mathbf{x})$ rather than $\sqrt{q}$. This point is significant because for many examples of structured signals that are encountered in practice, it is typically the case that $\eta_C(\mathbf{x}) \ll \sqrt{q}$. For example, if $\mathbf{x}$ is an s-sparse vector in $\mathbb{R}^q$ then proximal denoising via the ℓ_1-norm gives $\eta_{\ell_1}(\mathbf{x}) \leq \sqrt{2s\log(q/s) + 3s/2} + 7$ [35, 56, 121, 138]. Similarly, if $\mathbf{x}$ is a rank-r matrix in $\mathbb{R}^{p \times p}$ then proximal denoising via the matrix nuclear norm gives $\eta_{\mathrm{nuc}}(\mathbf{x}) \leq \sqrt{6rp} + 7$ [35, 56, 105, 117].

In order to state performance guarantees for a *sequence* of observations, we extend the definition of η_C to collections of vectors $\mathscr{X} = \{\mathbf{x}^\star[1], \ldots, \mathbf{x}^\star[n]\}, \mathbf{x}^\star[i] \in \mathbb{R}^q$ as

follows:

$$\eta_C(\mathscr{X}) := \inf_{\lambda \geq 0} \max_{\mathbf{x}^\star[t] \in \mathscr{X}} \left\{ \mathop{\mathbb{E}}_{\mathbf{g} \sim \mathcal{N}(0, I_{q \times q})} [\mathrm{dist}(\mathbf{g}, \lambda \cdot \partial \|\mathbf{x}^\star[t]\|_C)] \right\}. \tag{5.7}$$

5.3 Convex Programming for Change-Point Estimation

In this section, we describe our algorithm for high-dimensional change-point estimation by combining the filtered derivative method with proximal denoising. We state the main theorem that characterizes the accuracy of the estimated set of change-points, and we outline the proof, with the full details given in the Appendix.

5.3.1 Motivation

We begin by highlighting some of the difficulties that arise in change-point estimation as a result of the high-dimensionality of the observations. In order to frame our discussion concretely, we consider the prominent and widely-employed class of change-point estimation techniques based on the filtered derivative algorithm [7, 15, 17, 18], although similar difficulties arise with other approaches as well. The *filtered derivative* method detects changes based on an application of a pairwise difference operator to the output of a suitable low-pass filter applied to the sequence of observations. For simplicity, we describe a particular low-pass filter that is given by the sample mean of the observations over a small window (again, elaborations on this scheme are possible, with qualitatively similar conclusions). Formally, consider the following sequence defined at time t by computing differences of sample means over windows of size θ:

$$\mathrm{FD}_\theta[t] = -\frac{1}{\theta} \sum_{i=t-\theta+1}^{t} \mathbf{y}[i] + \frac{1}{\theta} \sum_{i=t+1}^{t+\theta} \mathbf{y}[i]. \tag{5.8}$$

Locations at which $\mathrm{FD}_\theta[t]$ has large magnitude (i.e., above a suitably chosen threshold) are declared as change-points.

This approach is well-suited for settings with sequences of scalar-valued signals, i.e., each $\mathbf{x}^\star[t]$ is scalar; see [7, 17, 18] for detailed analyses. However, if applied directly to the high-dimensional setting, the underlying sequence of signals $\mathbf{x}^\star[t] \in \mathbb{R}^q$ is required to remain stationary over time scales on the order of q so that changes can be reliably estimated. This requirement is unfortunately not realistic for practical purposes, e.g., in image processing applications one typically encounters $q \approx 10^6$. As such, it is desirable to develop an algorithm that detects changes in sequences of high-dimensional observations reliably even if the signal does not remain stationary over long time scales.

5.3.2 Our approach to high-dimensional change-point estimation

We base our method on the principle that more effective signal denoising by exploiting the low-dimensional structure underlying the sequence $\mathbf{x}^\star[t]$ enables improved change-point estimation. The formal steps of our algorithm for obtaining an estimate $\hat{\tau}$ of $\tau^\star$ are as follows:

1. **[Input]:** $\{\mathbf{y}[t]\}_{t=1}^n$ the sequence of signal observations, a choice of parameters θ, γ, λ to be employed in the algorithm, and a specification of a convex set C.

2. **[Filtering]:** Compute the moving averages $\bar{\mathbf{y}}[i] = \frac{1}{\theta}\sum_{t=i}^{i+\theta-1} \mathbf{y}[t], 1 \leq i \leq n-\theta+1$.

3. **[Denoising]:** Let $\hat{\mathbf{x}}[t], 1 \leq t \leq n-\theta+1$ be the solutions to the following convex optimization problems:

$$\hat{\mathbf{x}}[t] = \underset{\mathbf{x}\in\mathbb{R}^q}{\arg\min} \frac{1}{2}\|\bar{\mathbf{y}}[t] - \mathbf{x}\|_{\ell_2}^2 + \lambda\|\mathbf{x}\|_C. \tag{5.9}$$

4. **[Differencing]:** Compute $S[t] := \|\hat{\mathbf{x}}[t+1] - \hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2}$ for $\theta \leq t \leq n-\theta$.

5. **[Thresholding]:** For all t such that $S[t] < \gamma$, set $S[t] = 0$.

6. **[Output]:** Let $\{i_1, i_2, \ldots\} \subseteq \{\theta, \theta+1, \ldots, n-\theta\}$ be the indices of the nonzero entries of $S[t]$. Divide the set $\{i_1, i_2, \ldots\}$ into disjoint subsets $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_r, r \in \mathbb{Z}$, so that $i_{j+1} - i_j \leq \theta \Leftrightarrow i_j, i_{j+1} \in \mathcal{G}_k$ and $i_{j+1} - i_j > \theta \Leftrightarrow i_j \in \mathcal{G}_k, i_{j+1} \in \mathcal{G}_{k+1}$. The estimates $\hat{t}_i$ are given by $\hat{t}_i := \arg\max_{t\in\mathcal{G}_i} S[t]$, $1 \leq i \leq r$, and the output is $\hat{\tau} = \{\hat{t}_i\}$.

Observe that the proximal denoising step is applied before the differencing step. This particular integration of proximal denoising and the filtered derivative ensures that the differencing operator is applied to estimates $\bar{\mathbf{x}}[t]$ that are closer to the underlying signal $\mathbf{x}^\star[t]$ than the raw averages $\bar{\mathbf{y}}[t]$ (due to the favorable denoising properties of the proximal denoiser). As discussed in Theorem 5.3.1, this leads to improved change-point performance in comparison to a pure filtered derivative method. However, the analysis of our approach is complicated by the introduction of the proximal denoising step; we discuss this point in greater detail in Section 5.3.3.

The parameter θ determines the window over which we compute the sample mean, and it controls the resolution to which we estimate change-points. A larger value

of θ allows the algorithm to detect small changes, although if θ is chosen too large, multiple change-points may be mistaken for a single change-point. A smaller choice of θ increases the resolution of the change-point estimates, but small changes cannot be reliably detected. The parameter γ specifies the threshold for declaring changes, and it governs the size of the change-points that can be reliably estimated. A small choice of γ allows the algorithm to detect smaller changes, but it also increases the occurrence of false positives. Conversely, a larger value of γ reduces the number of false positives, but only those changes that are sufficiently large in magnitude may be detected by the algorithm (i.e., the number of false negatives may increase). In Theorem 5.3.1, we give precise guidelines for the choices of the parameters $(\theta, \gamma, \lambda)$ to guarantee reliable change-point estimation under suitable conditions via the method described above.

Theorem 5.3.1 *Consider a sequence of observations* $\mathbf{y}[t] = \mathbf{x}^\star[t] + \boldsymbol{\varepsilon}[t]$, $t = 1, \ldots, n$, *where each* $\mathbf{x}^\star[t] \in \mathbb{R}^q$ *and each* $\boldsymbol{\varepsilon}[t] \sim \mathcal{N}(0, \sigma^2 I_{q\times q})$ *independently. Let* $\tau^\star \subset \{1, \ldots, n\}$ *be such that* $t \in \tau^\star \Leftrightarrow \mathbf{x}^\star[t] \neq \mathbf{x}^\star[t+1]$, *let* $\Delta_{\min} = \min_{t\in\tau^\star} \|\mathbf{x}^\star[t] - \mathbf{x}^\star[t+1]\|_{\ell_2}$, *let* $T_{\min} = \min_{t_i, t_j \in \tau^\star, t_i \neq t_j} |t_i - t_j|$, *and let* $\mathscr{X} = \{\mathbf{x}^\star[1], \ldots, \mathbf{x}^\star[n]\}$. *Suppose* $\Delta_{\min}$ *and* $T_{\min}$ *satisfy*

$$\Delta_{\min}^2 T_{\min} \geq 64\sigma^2 \{\eta_C(\mathscr{X}) + r\sqrt{2\log n}\}^2 \tag{5.10}$$

for some $r > 1$ *and some convex set* C, *where* $\eta_C(\mathscr{X})$ *is as defined in* (5.7), *and* $\tau^\star \subset \{T_{\min}/4, \ldots, n - T_{\min}/4\}$. *Suppose we apply our change-point estimation algorithm with any choice of parameters* θ, γ, *and* λ *satisfying*

1. $T_{\min}/4 \geq \theta$,
2. $\Delta_{\min}/2 \geq \gamma \geq 2\frac{\sigma}{\sqrt{\theta}}\{\eta_C(\mathscr{X}) + r\sqrt{2\log n}\}$, *and*
3. $\lambda = \frac{\sigma}{\sqrt{\theta}} \arg\min_{\tilde{\lambda}} \max_{\mathbf{x}^\star[t]\in\mathscr{X}} \left\{ \mathop{\mathbb{E}}_{\mathbf{g}\sim\mathcal{N}(\mathbf{0}, I_{q\times q})} [\mathrm{dist}(\mathbf{g}, \tilde{\lambda} \cdot \partial \|\mathbf{x}^\star[t]\|_C)] \right\}$.

Then the algorithm recovers an estimate of the change-points $\hat{\tau}$ *satisfying*

1. $|\hat{\tau}| = |\tau^\star|$
2. $|\hat{t}_i - t_i^\star| \leq \min\{(4r\sqrt{\log n}/\eta_C(\mathscr{X}) + 4)\frac{\sigma\eta_C(\mathscr{X})}{\Delta\min}\sqrt{\theta}, \theta\}$ *for all* i, *where* $\hat{t}_i$ *and* $t_i^\star$ *are the* i*'th elements of* $\hat{\tau}$ *and* $\tau^\star$ *when ordered sequentially,*

with probability greater than $1 - 5n^{1-r^2}$.

Remarks. (1) If condition (5.10) is satisfied, then the choices of $\theta = T_{\min}/4$ and $\gamma = \Delta_{\min}/2$ satisfy the requirements in Theorem 5.3.1. Just as in the pure filtered derivative setting, a suitable choice of the parameter θ often relies on knowledge of the quantity $T_{\min}$ and is usually set at a constant times smaller than $T_{\min}$. In the setting where a desired total number of estimated change-points is available, the threshold γ may be set such that the output of our algorithm contains the desired number of change-points.

(2) For certain types of signal structure, one can specify suitable choices of λ that only depend on knowledge of the ambient dimension q. For example, in settings in which $\mathscr{X}$ is a collection of sparse vectors in $\mathbb{R}^q$, and we apply a proximal denoising step with the ℓ_1-norm, one may select $\lambda = \sigma\sqrt{2\log(q)/\theta}$, and in settings in which $\mathscr{X}$ is a collection of $p \times p$ low-rank matrices, and we apply a denoising step with the nuclear-norm, one may select $\lambda = 2\sigma\sqrt{p/\theta}$ [106].

(3) The performance of our algorithm is robust to misspecification of the convex set C. In particular, the quantity $\eta_C(\mathscr{X})$ for a misspecified C is in general larger than that for the correct C, but it is always smaller than $\sqrt{q}$. Consequently, the recovery guarantees associated with using a misspecified set C are weaker in general, but no worse than applying a pure filtered derivative algorithm without a denoising step.

(4) Our results can be extended to settings in which the noise $\boldsymbol{\varepsilon}[t]$ has correlations over space or time. For example, if the noise is distributed as $\boldsymbol{\varepsilon}[t] \sim \mathcal{N}(0, \Sigma)$, one could apply our algorithm to the transformed sequence of observations $\{\Sigma^{-1/2}\mathbf{y}[t]\}_{t=1}^n$ with the convex set C in the denoising step being replaced by the set $\Sigma^{-1/2}C$. If the noise $\boldsymbol{\varepsilon}[t]$ is correlated across time, one could apply a temporal whitening filter before proceeding with our algorithm. The application of such a filter generally leads to a smoothening of the abrupt changes that occur at change-points. As long as the bandwidth of the noise correlation is much smaller than $T_{\min}$, applying our algorithm to the smoothened sequence leads to qualitatively similar guarantees as in the case in which the noise is independent across time.

As a concrete illustration of this theorem, if each element $\mathbf{x}^\star[t] \in \mathbb{R}^q$, $t = 1, \ldots, n$ of the signal sequence is a vector consisting of at most s nonzero entries, then our algorithm (with a proximal denoiser based on the ℓ_1-norm) estimates change-points reliably under the condition $\Delta_{\min}^2 T_{\min} \gtrsim \sigma^2(s\log(\frac{q}{s}) + \log(n))$. Similarly, if each element $\mathbf{x}^\star[t] \in \mathbb{R}^{p\times p}$, $t = 1, \ldots, n$ is a matrix with rank at most r, then our algorithm

(with a proximal denoiser now based on the nuclear norm) provides reliable change-point estimation performance under the condition $\Delta_{\min}^2 T_{\min} \gtrsim \sigma^2(rp + \log(n))$.

5.3.3 Proof of Theorem 5.3.1

The proof broadly proceeds by bounding the probabilities of the following three events:

$$\mathcal{E}_1 := \{S[t] \geq \gamma, \forall t \in \tau^\star\} \tag{5.11}$$

$$\mathcal{E}_2 := \{S[t] < \gamma, \forall t \in \tau_{\text{far}}\} \tag{5.12}$$

$$\mathcal{E}_3 := \{\|\hat{\mathbf{x}}[t+1] - \hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2} > \|\hat{\mathbf{x}}[t+1+\delta] - \hat{\mathbf{x}}[t-\theta+1+\delta]\|_{\ell_2}, \forall (t,\delta) \in \tau_{\text{buffer}}\}. \tag{5.13}$$

Here $\tau_{\text{far}} = \{i : \theta \leq i \leq n-\theta, |i-j| > \theta, j \in \tau^\star\}$ and $\tau_{\text{buffer}} = \{(t_i^\star, \delta) : t_i^\star \in \tau^\star, \theta \geq |\delta| > (4r\sqrt{\log n}/\eta_C(\mathscr{X}) + 4)\frac{\sigma\eta_C(\mathscr{X})}{\Delta_{\min}}\sqrt{\theta}\}$. Note that τ_{buffer} defines a non-empty set if $\theta > (4r\sqrt{\log n}/\eta_C(\mathscr{X}) + 4)^2\sigma^2\eta_C^2(\mathscr{X})/\Delta_{\min}^2$. The event $\mathcal{E}_1$ corresponds to the atomic-norm-thresholded derivative exceeding the threshold γ for all change-points, while event $\mathcal{E}_2$ corresponds to the atomic-norm-thresholded derivative *not* exceeding the threshold γ in regions 'far away' from the change-points. Bounding the probabilities of these two events is sufficient for a weaker recovery guarantee than is provided by Theorem 5.3.1, which is that any estimated change-point $\hat{t} \in \hat{\tau}$ will be within θ of an actual change-point $t^\star \in \tau^\star$. However, the selection of the *maximum* derivative in Step 6 of the algorithm often leads to far more accurate estimates of the locations of change-points. To prove that this is the case, we consider the event $\mathcal{E}_3$ corresponding to the atomic-norm-thresholded derivative at the change-point being larger than the atomic-norm-thresholded derivatives at other points that are still within a window of θ but outside a small buffer region around the change-point.

The next proposition gives bounds on the probabilities of the events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$:

Proposition 5.3.2 *Under the setup and conditions of Theorem 5.3.1, we have the following bounds:*

$$\mathbb{P}(\mathcal{E}_1^c) \leq 2n^{1-r^2}, \quad \mathbb{P}(\mathcal{E}_2^c) \leq 2n^{1-r^2}, \quad \mathbb{P}(\mathcal{E}_3^c) \leq n^{1-r^2}. \tag{5.14}$$

The events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ are defined in (5.11)*,* (5.12)*, and* (5.13)*.*

The proof of Proposition 5.3.2 is given in the Appendix, and it involves overcoming two difficulties. First, if the filtering operator is applied over a window containing a

change-point in Step 2, the average $\bar{\mathbf{y}}[t]$ is in effect a superposition of two structured signals corrupted by noise. This necessitates the analysis of the performance of a proximal denoiser applied to a noisy superposition of structured signals rather than to a single structured signal corrupted by noise. The second (more challenging) complication arises due to the fact that the differencing operator is applied to the result of a nonlinear mapping of the observations (via the proximal denoiser) rather than to just a linear average of the observations as in a standard filtered derivative framework. We address these difficulties by exploiting certain properties of the proximal operator such as its non-expansiveness and its robustness to perturbations. Assuming Proposition 5.3.2, the proof of Theorem 5.3.1 proceeds as follows:

Proof of Theorem 5.3.1. From Proposition 5.3.2 and the union bound we have that $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \geq 1 - 5n^{r^2-1}$. We condition on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ to complete the proof. Specifically, conditioning on $\mathcal{E}_1$ ensures that we have $S[t] \geq \gamma$ for all $t \in \tau^\star$ after Step 5. Conditioning on $\mathcal{E}_2$ implies that all entries of S outside a window of θ from any change-point are set to 0 after Step 5. Hence, after Step 6 all non-zero entries of S that are within a window of at most θ around a change-point will have been grouped together (for all change-points), which implies that $|\hat{\tau}| = |\tau^\star|$. Finally, conditioning on the event $\mathcal{E}_3$ implies that $S[t]$ is larger than $S[t+\delta]$ for all δ such that $(4r\sqrt{\log n}/\eta_C(\mathscr{X}) + 4)\frac{\sigma\eta_C(\mathscr{X})}{\Delta\min}\sqrt{\theta} < |\delta| \leq \theta$ and for all $t \in \tau^\star$. Thus, our estimate of the change-point at $t \in \tau^\star$ is at most $\min\left\{(4r\sqrt{\log n}/\eta_C(\mathscr{X}) + 4)\frac{\sigma\eta_C(\mathscr{X})}{\Delta\min}\sqrt{\theta}, \theta\right\}$ away, which concludes the proof. □

5.3.4 Signal reconstruction

Based on the estimated set of change-points $\hat{\tau}$, it is straightforward to obtain good reconstructions of the signal away from the change-points via proximal denoising (5.5). This result follows from the analysis in [19, 106], but we state and prove it here for completeness:

Proposition 5.3.3 *Suppose that the assumptions of Theorem 5.3.1 hold. Let $t_1, t_2 \in \hat{\tau}$ be two consecutive estimates of change-points, let*

$$\lambda' = \frac{\sigma}{\sqrt{t_2 - t_1 - 2\theta}} \arg\min_{\tilde{\lambda}} \max_{\mathbf{x}^\star[t] \in \mathscr{X}} \{\mathbb{E}_{\mathbf{g} \sim \mathcal{N}(0, I_{q\times q})}[\mathrm{dist}(\mathbf{g}, \tilde{\lambda} \cdot \partial\|\mathbf{x}^\star[t]\|_C)]\},$$

and let $\bar{\mathbf{y}} = \frac{1}{t_2 - t_1 - 2\theta}\sum_{t=t_1+\theta+1}^{t_2-\theta} \mathbf{y}[t]$. Denote the solution of the proximal denoiser (5.5) applied to $\bar{\mathbf{y}}$ as $\bar{\mathbf{x}}$:

$$\bar{\mathbf{x}} := \arg\min_{\mathbf{x}} \frac{1}{2}\left\|\bar{\mathbf{y}} - \mathbf{x}\right\|_{\ell_2}^2 + \lambda'\|\mathbf{x}\|_C.$$

Letting $\bar{\mathbf{x}}$ *be the estimate of the signal* $\mathbf{x}^\star[t]$ *over the interval* $\{t_1+\theta+1,\ldots,t_2-\theta\}$*, we have that*

$$\|\bar{\mathbf{x}}-\mathbf{x}^\star[t]\|_{\ell_2}^2 \leq \frac{2\sigma^2}{t_2-t_1-2\theta}\{\eta_C(\mathscr{X})^2+s^2\}$$

with probability greater than $1-4n^{1-r^2}-\exp(-s^2/2)$*, for all* $t_1+\theta+1 \leq t \leq t_2-\theta$*.*

The proof of Proposition 5.3.3 is given in the Appendix. In order to obtain an accurate reconstruction of the underlying signal at some point in time, Proposition 5.3.3 requires that the duration of stationary of the signal around that time instance be long (in addition to the conditions of Theorem 5.3.1 being satisfied).

5.4 Tradeoffs in High-Dimensional Change-Point Estimation

Data analysis in practice involves a range of challenges beyond the high-dimensionality of the observations that motivated our development in this paper. For example, in change-point estimation in financial time series, one is typically faced with additional difficulties arising from the extremely rapid rate at which the data are acquired and the requirement that the data be processed in an 'online' fashion, i.e., the change-point estimation procedure must process the incoming data in 'real time.' In some settings, rapid variations in an underlying phenomenon trigger frequent changes in the sequence of observations, while in other cases small changes in a signal can be difficult to detect when severely corrupted by noise. In this section, we describe adaptations of the algorithm proposed in Section 5.3 to handle some of these challenges. Specifically, Theorem 5.3.1 suggests a number of performance tradeoffs that can be obtained in change-point estimation problems by employing suitable variations of our algorithm. We demonstrate the utility of these modifications in addressing some of the difficulties mentioned above, which highlights the versatility of our approach.

5.4.1 Change-point frequency and size tradeoffs

The appearance of the term $\Delta_{\min}^2 T_{\min}$ in (5.10) suggests an explicit relation between the minimum time span between changes, the minimum size of a change, and estimation accuracy. To illustrate the tradeoffs between $\Delta_{\min}$ and $T_{\min}$ clearly, we fix the complexity parameter $\eta_C(\mathscr{X})$ and the number of observations n in this discussion. As a consequence, the quantity $\Delta_{\min}^2 T_{\min}$ in (5.10) can be interpreted as a *resolution* on the types of changes that can be detected. Specifically, Theorem 5.3.1 guarantees reliable estimation of changes whenever $\Delta_{\min}^2 T_{\min}$ is sufficiently large, even if one of $\Delta_{\min}$ or $T_{\min}$ may be small. Previous works in the change-

point estimation literature have also demonstrated performance guarantees that are suggestive of tradeoffs similar to ours [57, 76]. In our setting, the quantities $\Delta_{\min}$ and $T_{\min}$ inform us about the choices of key parameters in the algorithm.

In settings in which $\Delta_{\min}$ is small, i.e., there are change-points where the size of the change is small, Theorem 5.3.1 guarantees that all the changes can be detected reliably so long as the distance between change-points ($T_{\min}$) is sufficiently large. In order to accomplish this, one is required to choose a sufficiently small threshold parameter γ (for Step 5 of the procedure) and a suitably large averaging window θ (for Step 2) with $\theta \lesssim T_{\min}$, in accordance with the requirements of Theorem 5.3.1. By smoothing over large windows of size θ and subsequently applying a proximal denoiser, even small-sized changes can be detected as long as the averaging window does not include multiple change-points (hence the condition that $\theta \lesssim T_{\min}$). The downside with choosing a large value for the parameter θ is that we do not resolve the locations of the change-points well; in particular, we estimate the locations of each of the change-points to within a resolution of about $\sqrt{\theta}$. However, detecting small-sized changes in a sequence corrupted by noise necessitates the computation of averages over large windows in Step 2 of our algorithm in order to distinguish genuine changes from spurious ones. Therefore, the low resolution to which we estimate the locations of change-points is the price to pay for estimating the number of change-points exactly in settings in which some of the changes may be small in size.

In a similar manner, if changes occur frequently in a signal sequence, i.e., the distance between change-points $T_{\min}$ is small, Theorem 5.3.1 guarantees that all the changes can be detected reliably if the size of each change $\Delta_{\min}$ is sufficiently large. In such cases, the averaging window parameter θ must be chosen to be sufficiently small while the threshold parameter γ must be appropriately large with $\gamma \lesssim \Delta_{\min}$, as prescribed in Theorem 5.3.1. The choice of a small value for θ ensures that we do not smooth the observation sequence over windows that contain multiple change-points in Step 2 of our method. However, this restriction of the averaging window size implies that the proximal denoiser in Step 3 is applied to the average of a small number of observations, which negatively impacts its performance. This limitation underlies the choice of a large value for the threshold parameter γ in Step 5 of the algorithm, which ensures that spurious changes resulting from denoising over small windows do not impact the performance of our algorithm. Consequently, the size of each change must be sufficiently large (as required by the condition that $\gamma \lesssim \Delta_{\min}$)

so that the change-points can be reliably estimated from a few noisy observations. Analogous to the discussion in the previous paragraph with $\Delta_{\min}$ being small and $T_{\min}$ being large, we also face resolution issues in settings in which $T_{\min}$ is small and $\Delta_{\min}$ is large. Specifically, the quality of the estimate of the underlying signal at time $\tilde{t}$ is governed by the duration of stationarity of $\mathbf{x}[t]$ around $\tilde{t}$ (as discussed in Proposition 5.3.3). As a result of $T_{\min}$ being small, the increased frequency of changes leads to poor estimates of the signal in between change-points.

5.4.2 Computational complexity and sample size tradeoffs

In change-point estimation tasks arising in many contemporary problem domains (e.g., financial time series), one is faced with a twin set of challenges: (*a*) the number of observations n may be quite large due to the increasingly higher frequencies at which data are acquired (this is in addition to the high dimensionality of each observation), and (*b*) the requirement that these large datasets be processed online or in real time. Consequently, as the number of observations per unit time grows, it is crucial that we adopt a *simpler* algorithmic strategy – i.e., a method requiring a smaller number of computational steps per observation – so that the overall computational complexity of our algorithm does not grow with the number of observations. In this section we describe a convex relaxation approach to adapt the algorithm described in Section 5.3 to achieve a tradeoff between the number of observations and the overall computational complexity of our procedure; in particular, we demonstrate that in certain change-point estimation problems one can even achieve an overall *reduction* in computational runtime as the number of observations grows. Our method is based on the ideas presented in [33, 35] in the context of statistical denoising and of linear inverse problems; here we demonstrate the utility of those insights in change-point estimation. We note that other researchers have also explored the idea of trading off computational resources and sample size in various inferential problems such as binary classifier learning [23, 44, 126, 128], sparse principal component analysis [6, 16, 86], model selection [1], and linear regression [129].

A modified change-point estimation algorithm. For ease of analysis and exposition, we consider a modification of our change-point estimation procedure from Section 5.3. Specifically, Step 6 of our algorithm is simplified so it only groups time indices corresponding to the nonzero entries of the thresholded derivative values, with consecutive time indices in a group at most θ apart (i.e., without further

choosing the maximum element from each group). Thus, our algorithm only produces windows that localize change-points instead of returning precise estimates of change-points. The reason for restricting our attention to such a simplification is that the additional operation of choosing the maximum element in Step 6 of the original algorithm leads to unnecessary complications that are not essential to the point of the discussion in this section. The performance analysis of this modified algorithm follows from Theorem 5.3.1, and we record this result next:

Corollary 5.4.1 *Under the same setup and conditions as in Theorem 5.3.1, suppose that we perform the modified change-point estimation algorithm – that is, Step 6 is simplified to only return groups of times indices, where consecutive time indices in a group are at most θ apart. Then we have with probability greater than $1 - 4n^{1-r^2}$ that (i) there are exactly $|\tau^\star|$ groups, and (ii) the j'th group $\mathcal{G}_j \subset \{\theta, \ldots, n-\theta\}$ is such that $|t_j^\star - \tilde{t}| \leq \theta$ for all $\tilde{t} \in \mathcal{G}_j$.*

In order to concretely illustrate tradeoffs between the number of observations and the overall computational runtime, we focus on the following stylized change-point estimation problem. Consider a continuous-time piecewise constant signal $\mathbf{x}^\star(T) \in \mathbb{R}^q$, $T \in (0, 1]$ defined as:

$$\mathbf{x}^\star(T) = \mathbf{x}^{\star(i)}, \quad T \in (T_i, T_{i+1}].$$

That is, the signal $\mathbf{x}^\star(T)$ takes on the value $\mathbf{x}^{\star(i)} \in \mathbb{R}^q$ identically for the entire time interval $T \in (T_i, T_{i+1}]$ for $i = 1, \ldots, k$. Here $i = 1, \ldots, k$ and the time indices $\{T_i\}_{i=1}^{k+1}$ are such that $0 = T_1 \leq \cdots \leq T_{k+1} = 1$. Suppose we have two collections of noisy observations obtained by sampling the signal $\mathbf{x}^\star(T)$ at equally-spaced points $\frac{1}{n}$ apart and $\frac{1}{kn}$ apart for some positive integers $k > 1$ and n (we assume that $n \gg \frac{1}{T_{i+1}-T_i}$ for all i):

$$\begin{aligned} \mathbf{y}^{(1)}[t] &= \mathbf{x}^\star\left(\frac{t}{n}\right) + \boldsymbol{\varepsilon}[t], \quad t = 1, \ldots, n \\ \mathbf{y}^{(2)}[t] &= \mathbf{x}^\star\left(\frac{t}{kn}\right) + \tilde{\boldsymbol{\varepsilon}}[t], \quad t = 1, \ldots, kn. \end{aligned}$$

Here $\boldsymbol{\varepsilon}[t], \tilde{\boldsymbol{\varepsilon}}[t] \sim \mathcal{N}(0, \sigma^2 I_{q\times q})$ are i.i.d. Gaussian noise vectors. In words, $\mathbf{y}^{(2)}[t]$ is a k-times more rapidly sampled version of $\mathbf{x}^\star(T)$ than $\mathbf{y}^{(1)}[t]$. As a result, the sequence $\mathbf{y}^{(1)}[t]$ consists of n observations and the sequence $\mathbf{y}^{(2)}[t]$ consists of kn observations. Consequently, it may seem that estimating the change-points in the sequence $\mathbf{y}^{(2)}[t]$ requires at least as many computational resources as the estimation

of change-points in $\mathbf{y}^{(1)}[t]$. However, when viewed from the prism of Corollary 5.4.1 and Theorem 5.3.1, the sequence $\mathbf{y}^{(2)}[t]$ is in some sense more favorable than $\mathbf{y}^{(1)}[t]$ for change-point estimation – specifically, if the minimum distance between successive change-points underlying the sequence $\mathbf{y}^{(1)}[t]$ is $T_{\min}$, then the minimum distance between successive change-points in $\mathbf{y}^{(2)}[t]$ is $kT_{\min}$, i.e., larger by a factor of k. (Note that $\Delta_{\min}$ for both sequences remains the same.)

Let $\mathscr{X} = \{\mathbf{x}^{\star(1)}, \ldots, \mathbf{x}^{\star(k)}\}$. Applying the modified change-point estimation algorithm described above to the sequence $\mathbf{y}^{(1)}[t]$ with parameters $\theta_1 = T_{\min}/4, \gamma_1 = \Delta_{\min}/2$ and with a proximal denoising step based on a convex set C, we obtain reliable localizations of the change-points under the condition:

$$\Delta_{\min}^2 T_{\min} \geq 64\sigma^2\{\eta_C(\mathscr{X}) + r\sqrt{2\log n}\}^2,$$

That is, we localize the change-points to a window of size $\theta_1 = T_{\min}/4$. Now suppose we apply the modified change-point algorithm to the sequence $\mathbf{y}^{(2)}[t]$ (note that this sequence is of length kn) with parameters $\theta_2 = kT_{\min}/4, \gamma_2 = \Delta_{\min}/2$ and with a proximal denoising step based on the same convex set C. In this case, we reliably localize each change-point in $\mathbf{y}^{(2)}[t]$ to a window of size $\theta_2 = kT_{\min}/4$ under the following condition:

$$\Delta_{\min}^2 (kT_{\min}) \geq 64\sigma^2\{\eta_C(\mathscr{X}) + r\sqrt{2\log n + 2\log k}\}^2, \tag{5.15}$$

The quality of the output in both cases is the same – identifying changes in $\mathbf{y}^{(1)}[t]$ to a resolution of $\theta_1 = T_{\min}/4$ is comparable to identifying changes in $\mathbf{y}^{(2)}[t]$ to a resolution of $\theta_2 = kT_{\min}/4$, because $\mathbf{y}^{(2)}[t]$ is a k-times more rapidly sampled version of the continuous-time signal $\mathbf{x}^{\star}(T)$ in comparison to $\mathbf{y}^{(1)}[t]$. On the computational side, our algorithm involves roughly n applications of the proximal denoiser based on the set C in the case of $\mathbf{y}^{(1)}[t]$, and about kn applications of the same proximal denoiser in the case of $\mathbf{y}^{(2)}[t]$. Therefore, the overall runtime is higher in the case of $\mathbf{y}^{(2)}[t]$ than in the case of $\mathbf{y}^{(1)}[t]$.

Notice that the left-hand side of the condition (5.15) goes up by a factor of k. We exploit this increased gap between the two sides of the inequality in (5.15) to obtain a smaller overall computational runtime for estimating changes in the sequence $\mathbf{y}^{(2)}[t]$ than for estimating changes in the sequence $\mathbf{y}^{(1)}[t]$. The key insight underlying our approach, borrowing from the ideas developed in [33, 35], is that we can employ a computationally cheaper proximal denoiser when applying our algorithm to the sequence $\mathbf{y}^{(2)}[t]$. Specifically, for many interesting classes of structured signals, one

can replace the proximal denoising operation with respect to the convex set C in Step 3 of our algorithm with a proximal operator corresponding to a *relaxation* $\mathcal{B} \subset \mathbb{R}^q$ of the set C, i.e., $\mathcal{B}$ is a convex set such that $C \subset \mathcal{B}$. For suitable relaxations $\mathcal{B}$ of the set C, the proximal denoiser associated to $\mathcal{B}$ is more efficient to compute than the proximal denoiser with respect to C, and further $\eta_C(\mathscr{X}) < \eta_{\mathcal{B}}(\mathscr{X})$. The reason for the second property is that, under appropriate conditions, the subdifferentials with respect to the gauge functions $\|\cdot\|_{\mathcal{B}}, \|\cdot\|_C$ satisfy the condition that $\partial\|\mathbf{x}^\star\|_{\mathcal{B}} \subset \partial\|\mathbf{x}^\star\|_C$ at signals of interest $\mathbf{x}^\star \in \mathscr{X}$. We refer the reader to [33] for further details, and more generally, to the convex optimization literature [69, 89, 110, 119, 130] for various constructions of families of tractable convex relaxations.

Going back to the sequence $\mathbf{y}^{(2)}[t]$, we can employ a proximal denoiser based on *any* tractable convex relaxation $\mathcal{B}$ of the set C as long as the following condition (a modification of (5.15)) for reliable change-point estimation is satisfied:

$$\Delta^2_{\min}(kT_{\min}) \geq 64\sigma^2\{\eta_{\mathcal{B}}(\mathscr{X}) + r\sqrt{2\log n + 2\log k}\}^2.$$

Indeed, if this condition is satisfied, we can still localize changes to a resolution of $kT_{\min}/4$, i.e., the same quality of performance as before with a proximal denoiser with respect to the set C. However, the computational upshot is that the number of operations required to estimate change-points in $\mathbf{y}^{(2)}[t]$ using the modified proximal denoising step is roughly kn applications of the proximal denoiser based on the relaxations $\mathcal{B}$. The contrast to n applications of a proximal denoiser based on C for estimating change-points in the sequence $\mathbf{y}^{(1)}[t]$ can be significant if the computation of the proximal denoiser with respect to $\mathcal{B}$ is much more tractable than the computation of the denoiser with respect to C.

We give an example in which such convex relaxations can lead to reduced computational runtime as the number of observations increases. We refer the reader to [33] for further illustrations in the context of statistical denoising, which can also be translated to provide interesting examples in a change-point estimation setting. Specifically, suppose that the underlying signal set $\mathscr{X} = \{\mathbf{a}\mathbf{a}' : \mathbf{a} \in \{-1, +1\}^p\}$, i.e., the signal at each instant in time is a rank-one matrix formed as an outer product of signed vectors. In this case, a natural candidate for a set C is the set of $p \times p$ *correlation matrices*, which is also called the *elliptope* in the convex optimization literature [46]. One can show that each application of a proximal denoiser with respect to C requires $O(p^{4.5})$ operations [78]. The $p \times p$ nuclear norm ball (scaled to contain all $p \times p$ matrices with nuclear norm at most p), which we denote as

$\mathcal{B}$, is a relaxation of the set C of correlation matrices. Interestingly, the distance $\eta_{\mathcal{B}}(\mathscr{X})$ is only a constant times larger (independent of the dimension p) than $\eta_C(\mathscr{X})$ [33]. However, each application of a proximal denoiser with respect to $\mathcal{B}$ requires only $O(p^3)$ operations. In summary, even if the increased sampling factor k in our setup is larger than a constant (independent of the dimension p), one can obtain an overall reduction in computational runtime from about $O(np^{4.5})$ operations to about $O(knp^3)$ operations.

5.5 Numerical Results

We illustrate the performance of our change-point estimation algorithm with two numerical experiments on synthetic data.

A contrast between our approach and the filtered derivative. The objective of the first experiment is to demonstrate the improved performance of our algorithm from Section 5.3.2 in comparison to the classical filtered derivative approach in a stylized problem setup. Recall that the filtered derivative method is equivalent to omitting the proximal denoising step in our algorithm, i.e., $\hat{\mathbf{x}}[t] = \bar{\mathbf{y}}[t]$ in Step 3 of our algorithm.

We consider a signal sequence $\mathbf{x}^\star[t] \in \mathbb{R}^{200\times 200}, t = 1,\dots,100$, consisting of exactly one change-point at time $t = 50$. Let $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)} \in \mathbb{R}^{200}$ be vectors with Euclidean-norm equal to 0.9 and direction chosen uniformly at random and independently. The signal $\mathbf{x}^\star[t]$ is a 200×200 matrix equal to $\mathbf{u}^{(1)}\mathbf{v}^{(1)\prime}$ before the change-point and $\mathbf{u}^{(2)}\mathbf{v}^{(2)\prime}$ after the change-point, and the observations are $\mathbf{y}[t] = \mathbf{x}^\star[t] + \boldsymbol{\varepsilon}[t], t = 1,\dots,100$, where $\boldsymbol{\varepsilon}[t] \sim \mathcal{N}(0, \sigma^2 I_{200^2\times 200^2})$ with $\sigma = 0.04$. Given this sequence of observations, we apply our algorithm with parameters $\lambda = 0.4$ and $\theta = 5$ (and with proximal denoising based on the nuclear-norm) and the filtered derivative algorithm with $\theta = 5$. The corresponding derivative values from our algorithm and the filtered derivative algorithm are given in the right sub-plot of Figure 5.1. We repeat the same experiment with the modification that the vectors $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)}$ now have Euclidean-norms equal to 2, thus leading to a larger-sized change relative to the noise. The corresponding derivative values from our algorithm and the filtered derivative algorithm are given in the right sub-plot in Figure 5.1.

One observation is that the derivative values are generally smaller with our approach than with the filtered derivative algorithm; this is primarily due to the inclusion of a denoising step, as a larger amount of noise leads to greater derivative values.

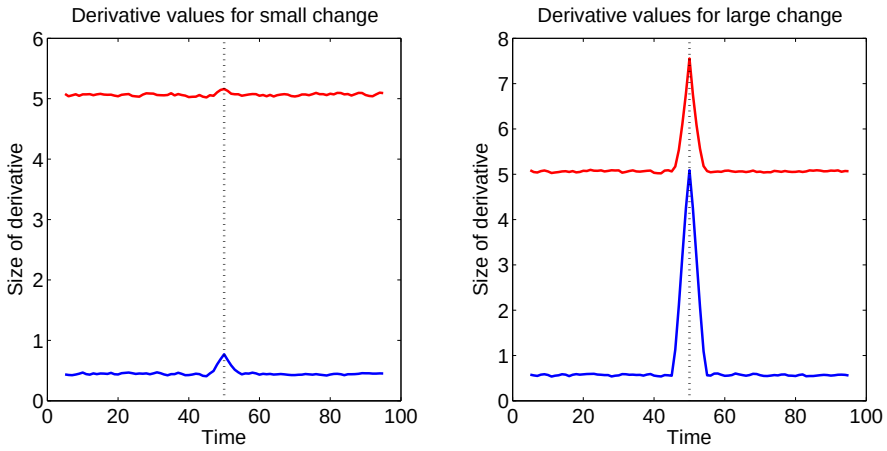


Figure 5.1: Experiment contrasting our algorithm (in blue) with the filtered derivative approach (in red): the left sub-plot corresponds to a small-sized change and the right sub-plot corresponds to a large-sized change.

More crucially, however, the *relative difference* in the derivative values near a change-point and away from a change-point is much larger with our algorithm than with the filtered derivative method. This is also a consequence of the inclusion of the proximal denoising step in our algorithm and the lack of a similar denoising operation in the filtered derivative approach. By suppressing the impact of noise via proximal denoising, our approach identifies smaller-sized changes more reliably than a standard filtered derivative method without a denoising step (see the sub-plot on the left in Figure 5.1).

Estimating change-points in sequences of sparse vectors. In our second experiment, we investigate the variation in the performance of our algorithm by choosing different sets of parameters θ, λ, γ. We consider a signal sequence $\mathbf{x}^\star[t] \in \mathbb{R}^{1000}$, $t = 1, \ldots, 1000$ consisting of sparse vectors. Specifically, we begin by generating 10 sparse vectors $\mathbf{s}^{(k)} \in \mathbb{R}^{1000}$, $k = 1, \ldots, 10$ as follows: for each $k = 1, \ldots, 10$, the vector $\mathbf{s}^{(k)}$ is a random sparse vector consisting of 30 nonzero entries (the locations of these entries are chosen uniformly at random and independently of k), with each nonzero entry being set to 1.2^{k-1}. We obtain the signal sequence $\mathbf{x}^\star[t]$ from the $\mathbf{s}^{(k)}$'s by setting $\mathbf{x}^\star[t] = \mathbf{s}^{(k)}$ for $t \in \{100(k-1)+1, 100k\}$. In words, the signal sequence $\mathbf{x}^\star[t]$ consists of 10 equally-sized blocks of length 100, and within each block the signal is identically equal to a sparse vector consisting of 30 nonzero entries. The magnitudes of the nonzero entries of $\mathbf{x}^\star[t]$ in the latter blocks are larger than those in the earlier blocks. The observations are $\mathbf{y}[t] = \mathbf{x}^\star[t] + \boldsymbol{\varepsilon}[t]$, $t = 1, \ldots, 1000$, where each $\boldsymbol{\varepsilon}[t] \sim \mathcal{N}(0, \sigma^2 I_{1000\times 1000})$ with σ chosen to be 2.5. We then apply our proposed

Run	θ	λ	γ
1	10	1	15
2	10	2	10
3	30	0.5	9
4	30	1	5

Table 5.1: Table of parameters employed in our change-point estimation algorithm in synthetic experiment with sparse vectors.

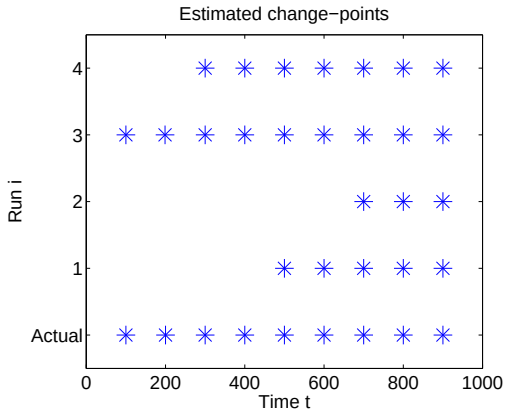


Figure 5.2: Plot of estimated change-points: the locations of the actual change-points are indicated in the bottom row.

algorithm using the four choices of parameters listed in Figure 5.1, with a proximal operator based on the ℓ_1-norm. The estimated sets of change-points are given in Figure 5.2, and the derivative values corresponding to Step 4 of our algorithm are given in Figure 5.3.

First, note that the algorithm generally detects smaller sized changes with larger values of θ and smaller values of γ (corresponding to Runs 3 and 4 from Figure 5.1), i.e., the averaging window size is larger in Step 2 of our algorithm, and the threshold is smaller in Step 4. Next, consider the graph of derivative values in Figure 5.3. The estimated locations of change-points correspond to peaks in Figure 5.3, so the algorithm can be interpreted as selecting a subset of peaks that are sufficiently separated (Step 6). We note that a smaller choice of θ leads to sharper peaks (and hence, smaller-sized groups in Step 6), while a larger choice of θ leads to wider peaks (correspondingly, larger-size groups in Step 6).

Phase transition. In the third experiment, we examine the performance of our algorithm for signal sequences with different values of $\Delta_{\min}$ and $T_{\min}$. For each $\Delta_{\min} \in \{\sqrt{4}, \sqrt{8}, \ldots, \sqrt{80}\}, T_{\min} \in \{4, 8, \ldots, 80\}$, we construct a sequence $\mathbf{x}^\star[t] \in \mathbb{R}^{100}$, $t = 1, \ldots, 1000$ such that the size (in Euclidean norm) of each change equals

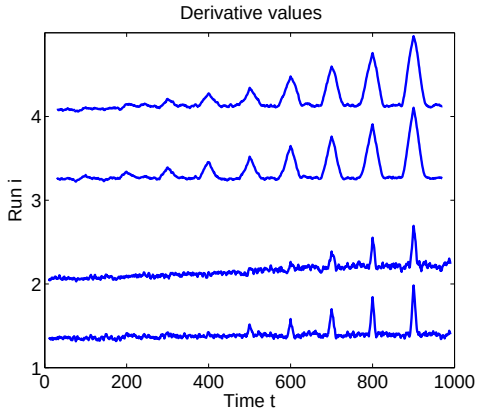


Figure 5.3: Experiment with sparse vectors: graphs of derivative values corresponding to different parameters choices from Figure 5.1.

$\Delta_{\min}$, and the interval between each consecutive pair of change-points is equal to $T_{\min}$. Specifically, we generate $\lfloor 1000/T_{\min} \rfloor$ sparse vectors $\mathbf{p}^{(k)} \in \mathbb{R}^{100}, 1 \leq k \leq \lfloor 1000/T_{\min} \rfloor$. The first 10 entries of the vector $\mathbf{p}^{(1)}$ are set to $\Delta_{\min}/\sqrt{20}$ and the remaining are set to zero. For each subsequent $\mathbf{p}^{(k)} \in \mathbb{R}^{100}$, $2 \leq k \leq \lfloor 1000/T_{\min} \rfloor$, we proceed sequentially by choosing 10 coordinates uniformly at random from those 90 coordinates of $\mathbf{p}^{(k-1)}$ that consist of zeros; we set these 10 coordinates of $\mathbf{p}^{(k)}$ to $\Delta_{\min}/\sqrt{20}$ and the remaining to zero. We obtain the signal sequence $\mathbf{x}^\star[t] \in \mathbb{R}^{100}$, $t = 1, \ldots, 1000$ by setting $\mathbf{x}^\star[t] = \mathbf{p}^{(k)}$ for $t \in \{(k-1)T_{\min} + 1, kT_{\min}\}$. The observations are given by $\mathbf{y}[t] = \mathbf{x}^\star[t] + \boldsymbol{\varepsilon}[t]$, $t = 1, \ldots, 1000$, with each $\boldsymbol{\varepsilon}[t] \sim \mathcal{N}(0, \sigma^2 I_{1000\times 1000})$ and $\sigma = 0.5$. We apply our algorithm from Section 5.3.2 with a proximal operator based on the ℓ_1-norm and with the choice of parameters $\lambda = \sigma\sqrt{2\log q} = \sqrt{\log 10}$ (in this example, $q = 100$), $\gamma = \Delta_{\min}/2$, and $\theta = T_{\min}/4$. For each $\Delta_{\min} \in \{\sqrt{4}, \sqrt{8}, \ldots, \sqrt{80}\}, T_{\min} \in \{4, 8, \ldots, 80\}$, we repeat this experiment 100 times.

We consider a trial to be a success if the two conclusions of Theorem 5.3.1 are achieved. First, the number of estimated change-points equals the true number of change-points. Second, each change-point estimate is within a window of size $\min\{(4r\sqrt{\log n}/\eta + 4)\frac{\sigma\eta}{\Delta\min}\sqrt{\theta}, \theta\}$ of an actual change-point, with $r = 1.2$, $n = 1000$, and $\eta = \sqrt{2s\log(q/s) + 1.5s} + 7$ is the Gaussian distance of s-sparse vectors in $\mathbb{R}^q$ from the discussion in Section 5.2.3 (in this example, $s = 10, q = 100$). Figure 5.4 shows the fraction of successful trials for different values of $\Delta_{\min}, T_{\min}$.

Observe that the frequency of successful trials is high when $\Delta_{\min}^2 T_{\min}$ is large, and that we see a phase transition in the performance of our approach as $\Delta_{\min}^2 T_{\min}$ becomes

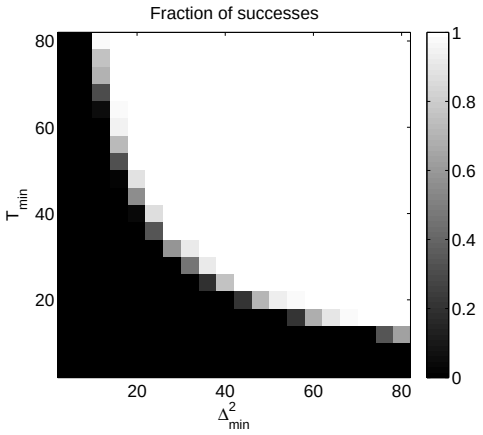


Figure 5.4: Experiment from Section 5.5 demonstrating a phase transition in the recovery of the set of change-points for different values of $\Delta_{\min}$ and $T_{\min}$. The black cells correspond to a probability of recovery of 0 and the white cells to a probability of recovery of 1.

small. In particular, the boundary of the phase transition appears to correspond to the quantity $\Delta_{\min}^2 T_{\min}$ being constant, which is the scaling law suggested by the recovery guarantee (5.10).

5.6 Conclusions

We propose an algorithm for high-dimensional change-point estimation that blends the filtered derivative method with a convex optimization step that exploits low-dimensional structure in the underlying signal sequence. We prove that our algorithm reliably estimates change-points provided the product of the square of the size of the smallest change (measured in ℓ_2-norm) and the smallest distance between changes is larger than Gaussian distance/width quantity η^2 that characterizes the low-dimensional complexity in the signal sequence. The dependence on η^2 is a result of the integration of the convex optimization step (based on proximal denoising).

The change-point literature also consists of extensive investigations of *quickest change detection* problems [94, 108, 114, 131, 149], which are qualitatively somewhat different than the setup considered in our work. In those settings one is given access to observations sequentially, and the objective is to correctly declare when a change-point occurs in the shortest time possible (i.e., minimize the expected delay) subject to false alarm rate constraints. Building on the algorithmic ideas described in this paper, it would be of interest to design computationally and statistically efficient techniques for high-dimensional quickest change detection problems by exploiting structure in the underlying signal sequence.

Chapter 6

CONCLUSIONS

We summarize the main contributions of this thesis, and discuss future research directions.

6.1 Summary of Contributions

Learning Semidefinite Regularizers

Regularizers are penalty functions that are deployed in optimization-based approaches for solving ill-posed inverse problems. The purpose of these functions is to induce a desired structure in the solution. In Chapter 2 we describe a framework for learning regularizers directly from data that can be computed efficiently using semidefinite programming. Our procedure for learning regularizers is based on computing a structured factorization of the data matrix. We develop an algorithm for computing such factorizations, and we prove local linear convergence of our method. Our numerical experiments on image denoising demonstrate the utility of our framework.

Fitting Tractable Convex Sets to Support Function Evaluations

The support function of a set gives the displacement of a supporting hyperplane from the origin. In Chapter 3 we describe a framework for fitting tractable convex sets to noisy support function evaluations. We prove statistical guarantees for our method. Our approach provides superior reconstructions in comparison to the previous approaches, most notably, in settings in which the measurements available are noisy or small in number as well as those in which the underlying set to be reconstructed is non-polyhedral.

Optimal Approximations of Convex Sets as Spectrahedra

In Chapter 4 we consider the problem of computing optimal approximations of convex sets as spectrahedra of a fixed size. We describe a numerical procedure for computing such approximations. Our techniques are useful for further understanding the expressiveness of semidefinite representations.

High-Dimensional Change-Point Estimation: Combining Filtering with Convex Optimization

In Chapter 5 we propose an algorithm that is suitable for estimating changes in sequences of high-dimensional time series. Our method combines a convex optimization-based step that exploits low-dimensional structure in the underlying signal sequence with a classical change-point estimation algorithm known as the filtered derivative. We provide performance guarantees of our algorithm, which exhibit a dependence on the low-dimensional complexity in the signal sequence.

6.2 Future Directions

Qualitative analysis of local minima in matrix factorization instances

In this thesis, we describe a framework for fitting convex sets to data through solving a non-convex optimization instance. The output of our algorithms for solving these instances depends on the choice of initialization. In practical settings, we observe that our methods are effective even though our algorithms typically do not converge to global minima. An important future direction is to explain why such behavior occurs.

One approach is to analyze the properties of local minima. In the context of learning a regularizer from data as we did in Chapter 2, it would be interesting to show that most local minima capture the 'correct' facial geometry. Such a result implies that these minima define regularizers that are equally effective.

Providing progress in this direction is important for a number of reasons. First, we show that the utility of our procedures is not contingent on computing the global minimizer. Second, we provide concrete a explanation as to why random initializations are useful in practice.

Understanding the expressiveness and utility of semidefinite descriptions

It is of deep interest to further understand the expressiveness of semidefinite descriptions. It is also of deep interest to identify settings in which these descriptions are useful in comparison to polyhedral descriptions.

A potential direction in the context of learning data representations is as follows. In many applications, it is natural to consider bases that are invariant under certain group transformations, and in particular, those that reflect physical considerations. For instance, we may be interested in learning a basis for natural images that is invariant to scalings, rotations, and translations. It is of interest to understand if such bases are naturally amenable to descriptions based on semidefinite programming.

Optimizing over affine slices of structured convex cones

Our numerical algorithms for fitting a convex set to data are based on searching over the space of projection of a *fixed* set that can be expressed as the intersection of a structured convex cone and an affine subspace. A richer family of convex sets are those that are expressible as the projection of the intersection of a structured convex cone and *any* affine subspace. An interesting direction is to develop numerical procedures for fitting convex sets to data over such a parameterization; i.e., we optimize over the space of linear projection maps and affine subspaces simultaneously.

Appendix A

PROOFS FOR CHAPTER 2

A.1 Proofs of Lemma 2.3.9 and Lemma 2.3.10

Proof of Lemma 2.3.9. Note that if $Z \in \mathcal{T}$ then Z has rank at most $2r$. As a consequence of the restricted isometry property we have $(1-\delta_{2r})\|Z\|_{\ell_2}^2 \leq \|[\mathcal{L} \circ \mathcal{P}_{\mathcal{T}}](Z)\|_{\ell_2}^2 \leq (1+\delta_{2r})\|Z\|_{\ell_2}^2$. Since $Z \in \mathcal{T}$ is arbitrary we have $1-\delta_{2r} \leq \lambda(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}}) \leq 1+\delta_{2r}$, which proves (i). This immediately implies the bound in (ii). Moreover since $\|\mathcal{L}\circ\mathcal{P}_{\mathcal{T}}\|_2 = \|\mathcal{P}_{\mathcal{T}}\circ\mathcal{L}'\mathcal{L}\circ\mathcal{P}_{\mathcal{T}}\|_2^{1/2} \leq \sqrt{1+\delta_{2r}}$, we have $\|\mathcal{P}_{\mathcal{T}}\circ\mathcal{L}'\mathcal{L}\|_2 \leq \sqrt{1+\delta_{2r}}\|\mathcal{L}\|_2$, which is (iii). Last we have $\|[(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}'\mathcal{L}\|_2 \leq \|[(\mathcal{L}'_{\mathcal{T}}\mathcal{L}_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}}\|_2\|\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}\|_2 \leq \frac{\sqrt{1+\delta_{2r}}}{1-\delta_{2r}}\|\mathcal{L}\|_2$, which proves (iv). □

Proof of Lemma 2.3.10. Since $\mathrm{tr}(\mathrm{cov}(\mathscr{X})) = \frac{1}{n}\sum_{j=1}^n \|X^{(j)}\|_{\ell_2}^2$, we have $s_{\min} \leq \mathrm{tr}(\mathrm{cov}(\mathscr{X})) \leq s_{\max}$. Next we have the inequalities $(\Lambda(\mathscr{X})-\Delta(\mathscr{X}))\mathrm{I} \leq \mathrm{cov}(\mathscr{X}) \leq (\Lambda(\mathscr{X})+\Delta(\mathscr{X}))\mathrm{I}$. The result follows by applying trace. □

A.2 Proof of Proposition 2.3.8

In this section we prove that the ensemble of random matrices $\mathscr{X}$ described in Proposition 2.3.8 satisfies the deterministic conditions in Theorem 2.3.5 with high probability. We begin with computing $\mathbb{E}_{\mathcal{D}}[X^{(j)} \boxtimes X^{(j)}]$, and $\mathbb{E}_{\mathcal{D}}[(X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})}]$. Note that the random matrices $\{X^{(j)} \boxtimes X^{(j)}\}_{j=1}^n$ and the random operators $\{(X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})}\}_{j=1}^n$ are almost surely bounded above in spectral norm by construction. This allows us to conclude Proposition 2.3.8 with an application of the Matrix Hoeffding Inequality [146].

To simplify notation we adopt the following. In the first two results we omit the superscript j from $X^{(j)}$. In the remainder of the section we let $\mathbb{E} = \mathbb{E}_{\mathcal{D}}$, $\bar{s}^2 := \mathbb{E}[s^2]$, $\{\mathbf{e}_i\}_{i=1}^p \subset \mathbb{R}^q$ be the set of standard basis vectors, and $\{E_{ij}\}_{i,j=1}^p \subset \mathbb{R}^{p\times p}$ be the set of matrices whose (i,j)-th entry is 1 and is 0 everywhere else.

Proposition A.2.1 *Suppose $X \sim \mathcal{D}$ as described in Proposition 2.3.8. Then $\mathbb{E}[X \boxtimes X] = \bar{s}^2(r/p^2)\mathrm{I}$.*

Proof. It suffices to show that $\mathbb{E}\langle X \boxtimes X, \mathbf{e}_w\mathbf{e}'_x \boxtimes \mathbf{e}_y\mathbf{e}'_z\rangle = \mathbb{E}\langle X, \mathbf{e}_w\mathbf{e}'_x\rangle\langle X, \mathbf{e}_y\mathbf{e}'_z\rangle = \delta_{wy}\delta_{xz}\bar{s}^2(r/p^2)$. Let $X = \sum_{i=1}^r s_i\mathbf{u}_i\mathbf{v}'_i$ as described in the statement of Proposition

2.3.8. Suppose we denote $\mathbf{u}_i = (u_{i1}, \ldots, u_{iq})'$, and $\mathbf{v}_i = (v_{i1}, \ldots, v_{iq})'$. By applying independence we have $\mathbb{E}\langle X, \mathbf{e}_w\mathbf{e}_x'\rangle\langle X, \mathbf{e}_y\mathbf{e}_z'\rangle = \mathbb{E}[(\sum_{i=1}^r s_i u_{iw} v_{ix})(\sum_{k=1}^r s_k u_{ky} v_{kz})] = \sum_{i,k=1}^r \mathbb{E}[s_i s_k]\mathbb{E}[u_{iw}u_{ky}]\mathbb{E}[v_{ix}v_{kz}]$. There are two cases we need to consider.

[Case $w \neq y$ or $x \neq z$]: Without loss of generality suppose that $w \neq y$. Then $\mathbb{E}[u_{iw}u_{ky}] = 0$ for all $1 \leq i, k \leq q$, and hence $\mathbb{E}\langle X \boxtimes X, \mathbf{e}_w\mathbf{e}_x' \boxtimes \mathbf{e}_y\mathbf{e}_z'\rangle = 0$.

[Case $w = y$ and $x = z$]: Note that if $i \neq k$ then $\mathbb{E}[u_{iw}u_{ky}] = \mathbb{E}[u_{iw}]\mathbb{E}[u_{ky}] = 0$. Since $\mathbf{u}_i$ is a unit-norm vector distributed u.a.r., we have $\mathbb{E}[u_{ix}^2] = 1/q$. Hence $\mathbb{E}\langle X \boxtimes X, \mathbf{e}_w\mathbf{e}_x' \boxtimes \mathbf{e}_y\mathbf{e}_z'\rangle = \sum_{i=1}^r \mathbb{E}[s_i^2]\mathbb{E}[u_{iw}^2]\mathbb{E}[v_{ix}^2] = \bar{s}^2 r/p^2$. □

Our next result requires the definition of certain subspaces of $\mathbb{R}^{p\times p}$ and $\mathrm{End}(\mathbb{R}^{p\times p})$.

We define the following subspaces in $\mathbb{R}^{p\times p}$: Let $\mathcal{G} := \{W : W = W', W \in I^\perp\}$ be the subspace of symmetric matrices that are orthogonal to the identity, $\mathcal{H} := \{W : W = -W'\}$ be the subspace of skew-symmetric matrices, and $\mathcal{I} = \mathrm{Span}(I)$. It is clear that $\mathbb{R}^{p\times p} = \mathcal{G} \oplus \mathcal{H} \oplus \mathcal{I}$.

In addition to the subspace $\mathcal{W}$ defined in (2.12), we define the following subspaces in $\mathrm{End}(\mathbb{R}^{p\times p})$:

1. $\mathcal{W}_{SS} := \mathrm{Span}(\{A \otimes B : A, B \in \mathcal{G}\})$,
2. $\mathcal{W}_{AA} := \mathrm{Span}(\{A \otimes B : A, B \in \mathcal{H}\})$,
3. $\mathcal{W}_{SA} := \mathrm{Span}(\{A \otimes B : A \in \mathcal{G}, B \in \mathcal{H}\})$,
4. $\mathcal{W}_{AS} := \mathrm{Span}(\{A \otimes B : A \in \mathcal{H}, B \in \mathcal{G}\})$.

Note that $\mathrm{End}(\mathbb{R}^{p\times p}) = \mathcal{W} \oplus \mathcal{W}_{SS} \oplus \mathcal{W}_{AA} \oplus \mathcal{W}_{SA} \oplus \mathcal{W}_{AS}$. To verify this, first express an arbitrary linear map $\mathsf{E} \in \mathrm{End}(\mathbb{R}^{p\times p})$ as a sum of Kronecker products $\mathsf{E} = \sum_{i=1} A_i \otimes B_i$, second decompose each matrix A_i, B_i into components in the subspaces $\{\mathcal{G}, \mathcal{H}, \mathcal{I}\}$, and third expand the expression. The orthogonality between subspaces is immediate from the identity $\langle A_i \otimes B_i, A_j \otimes B_j\rangle = \langle A_i, A_j\rangle\langle B_i, B_j\rangle$.

Proposition A.2.2 *Suppose $X \sim \mathcal{D}$ as described in Proposition 2.3.8. Then*

$$\mathbb{E}[(X \boxtimes X) \otimes \mathcal{P}_{\mathcal{T}(X)}] = c_{\mathcal{W}}\mathsf{I}_{\mathcal{W}} + c_{\mathcal{W}_{SS}}\mathsf{I}_{\mathcal{W}_{SS}} + c_{\mathcal{W}_{AA}}\mathsf{I}_{\mathcal{W}_{AA}} + c_{\mathcal{W}_{SA}}\mathsf{I}_{\mathcal{W}_{SA}} + c_{\mathcal{W}_{AS}}\mathsf{I}_{\mathcal{W}_{AS}},$$

where (i) $c_{\mathcal{W}} = \bar{s}^2 r(\frac{1}{p^2})$, (ii) $c_{\mathcal{W}_{SS}} = \bar{s}^2 r(\frac{1}{p^2} - \frac{(q-r)^2}{(q-1)^2(q+2)^2})$, (iii) $c_{\mathcal{W}_{AA}} = \bar{s}^2 r(\frac{1}{p^2} - \frac{(q-r)^2}{p^2(q-1)^2})$, and (iv) $c_{\mathcal{W}_{SA}} = c_{\mathcal{W}_{AS}} = \bar{s}^2 r(\frac{1}{p^2} - \frac{(q-r)^2}{q(q-1)^2(q+2)})$.

Proof. The proof consists of two parts, namely (i) to prove that the mean, when restricted to the respective subspaces described above, has diagonal entries as specified, and (ii) to prove that the off-diagonal elements are zero with respect to any basis that obeys the specified decomposition of $\mathrm{End}(\mathbb{R}^{p\times p})$. In addition, it suffices to only consider linear maps that are Kronecker products since these maps generate the respective subspaces. The following identity for all matrices A_i, B_i, A_j, B_j is particularly useful

$$\langle (A_i' \otimes B_i) \boxtimes (A_j' \otimes B_j), \mathbb{E}[(X \boxtimes X) \otimes \mathcal{P}_{\mathcal{T}(X)}] \rangle = \mathbb{E}\langle \mathcal{P}_{\mathcal{T}(X)}(B_j X A_j), \mathcal{P}_{\mathcal{T}(X)}(B_i X A_i) \rangle. \tag{A.1}$$

One may equivalently describe the distribution of X as follows – let $X = U\Sigma_R V'$, where U, V are $p \times p$ matrices drawn from the Haar measure, and Σ_R is a diagonal matrix whose first r entries are drawn from $\mathcal{D}$, and the remaining entries are 0 (to simplify notation we omit the dependence on X in the matrices U, V). Let $I_N = \mathrm{diag}(0, \ldots, 0, 1, \ldots, 1)$ be a diagonal matrix consisting of $p - r$ ones. Under this notation, the projector is simply the map $\mathcal{P}_{\mathcal{T}(X)}(Z) = Z - UI_N U' Z V I_N V'$. The remainder of the proof is divided into the two parts outlined above.

[Part (i)]: The restriction to diagonal entries correspond to the case $i = j$, and hence equation (A.1) simplifies to $\mathbb{E}[\|\mathcal{P}_{\mathcal{T}(X)}(BXA)\|_{\ell_2}^2]$. Consequently we have

$$\mathbb{E}[\|\mathcal{P}_{\mathcal{T}(X)}(BXA)\|_{\ell_2}^2] = \mathbb{E}[\|BU\Sigma_R V' A\|_{\ell_2}^2] - \mathbb{E}[\|I_N U' AU\Sigma_R V' BV I_N\|_{\ell_2}^2].$$

First we compute $\mathbb{E}[\|I_N U' AU\Sigma_R V' BV I_N\|_{\ell_2}^2]$. By the cyclicity of trace and iterated expectations we have

$$\begin{aligned} & \mathbb{E}[\|I_N U' AU\Sigma_R V' BV I_N\|_{\ell_2}^2] \\ = \; & \mathbb{E}[\mathrm{tr}(\Sigma_R^{1/2} U' A' U I_N U' AU\Sigma_R V' BV I_N V' B' V \Sigma_R^{1/2})] \\ = \; & \mathbb{E}_U[\mathbb{E}_V[\mathrm{tr}(\Sigma_R^{1/2} U' A' U I_N U' AU\Sigma_R V' BV I_N V' B' V \Sigma_R^{1/2})]]. \end{aligned}$$

It suffices to compute $\mathbb{E}[\Sigma_R^{1/2} V' BV I_N V' B' V \Sigma_R^{1/2}] = \Sigma_R^{1/2} \mathbb{E}[V' BV I_N V' B' V] \Sigma_R^{1/2}$ in the three cases corresponding to $B \in \{\mathcal{G}, \mathcal{H}, \mathcal{I}\}$ respectively. Using linearity and symmetry, it suffices to compute $\mathbb{E}[V' BV E_{11} V' B' V]$. We split this computation into the following three separate cases.

[Case $B \in \mathcal{I}$]: We have $I_N \Sigma_R^{1/2} = 0$, and hence the mean is the zero-matrix.

[Case $B \in \mathcal{H}$]: Claim: If $B \in \mathcal{H}$, and $\|B\|_{\ell_2} = 1$, then $\mathbb{E}[V' BV E_{11} V' B' V] = (I - E_{11})/(q(q-1))$.

Proof: Denote $V = [\mathbf{v}_1 | \ldots | \mathbf{v}_q]$. The off-diagonal entries vanish as $\mathbb{E}\langle E_{ij}, V'BVE_{11}V'B'V\rangle$ $= \mathbb{E}(\mathbf{v}_1'B\mathbf{v}_i)(\mathbf{v}_1'B\mathbf{v}_j) = 0$ whenever $i \neq j$, as one of the indices i, j appears exactly once. By a symmetry argument we have $\mathbb{E}[V'BVE_{11}V'B'V] = \alpha I + \beta E_{11}$ for some α, β. First $\mathbb{E}[\mathrm{tr}(V'BVE_{11}V'B'V)] = \mathbb{E}[\mathrm{tr}(BVE_{11}V'B')] = \mathrm{tr}(B\mathbb{E}[VE_{11}V']B') = \mathrm{tr}(B(I/q)B') = 1/q$, which gives $\alpha q + \beta = 1/q$. Second since B is asymmetric, $V'BV$ is also asymmetric and hence is 0 on the diagonals. Thus $\langle V'BVE_{11}V'B'V, E_{11}\rangle = 0$, which gives $\alpha + \beta = 0$. The two equations yield the values of α and β.

[Case: $B \in \mathcal{G}$]: Claim: If $B \in \mathcal{G}$, and $\|B\|_{\ell_2} = 1$, then $\mathbb{E}[V'BVE_{11}V'B'V] = (I + (1 - 2/q)E_{11})/((q-1)(q+2))$.

Proof: With an identical argument as the previous claim one has $\mathbb{E}[V'BVE_{11}V'B'V] = \alpha I + \beta E_{11}$, where $\alpha q + \beta = 1/q$. Next $\mathbb{E}[\langle V'BVE_{11}V'B'V, E_{11}\rangle] = \mathbb{E}[(\mathbf{v}_1'B\mathbf{v}_1)^2]$, where $\mathbf{v}_1$ is a unit-norm vector distributed u.a.r. Since conjugation by orthogonal matrices preserves trace, and $\mathbf{v}_1$ has the same distribution as $Q\mathbf{v}_1$ for any orthogonal Q, we may assume that $B = \mathrm{diag}(b_{11}, \ldots, b_{qq})$ is diagonal without loss of generality. Suppose we let $\mathbf{v}_1 = (v_1, \ldots, v_q)'$. Then $\mathbb{E}[(\mathbf{v}_1'B\mathbf{v}_1)^2] = \mathbb{E}[\sum b_{ii}^2 v_i^4 + \sum_{i\neq j} b_{ii}b_{jj}v_i^2v_j^2] = \mu_1(\sum b_{ii}^2) + \mu_2(\sum_{i\neq j} b_{ii}b_{jj})$, where $\mu_1 = \mathbb{E}[v_1^4]$, and $\mu_2 = \mathbb{E}[v_1^2v_2^2]$. Since $\mathrm{tr}(B) = 0$, we have $\sum b_{ii}^2 = -\sum_{i\neq j} b_{ii}b_{jj}$. Last from Theorem 2 of [37] we have $\mu_1 = 3/(q(q+2))$, and $\mu_2 = 1/(q(q+2))$, which gives $\mathbb{E}[(\mathbf{v}_1'B\mathbf{v}_1)^2] = 2/(q(q+2))$, and hence $\alpha + \beta = 2/(q(q+2))$. The two equations yield the values of α and β.

With a similar set of computations one can show that $\mathbb{E}[\|BU\Sigma_R V'A\|_{\ell_2}^2] = \bar{s}^2 r/p^2$ for arbitrary unit-norm A, B. An additional set of computations yields the diagonal entries, which completes the proof. We omit these computations.

[Part (ii)]: We claim that it suffices to show that $\mathbb{E}[V'A_iVE_{11}V'A_j'V]$ is the zero-matrix whenever $A_i, A_j \in \{\mathcal{G}, \mathcal{H}, \mathcal{I}\}$, and satisfy $\langle A_i, A_j\rangle = 0$. We show how this proves the result. Suppose $A_i \otimes B_i, A_j \otimes B_j$ satisfy $\langle A_i \otimes B_i, A_j \otimes B_j\rangle = \langle A_i, A_j\rangle\langle B_i, B_j\rangle = 0$. Without loss of generality we may assume that $\langle A_i, A_j\rangle = 0$. From equation (A.1) we have

$$\begin{aligned} & \mathbb{E}\langle \mathcal{P}_{\mathcal{T}(X)}(B_jXA_j), \mathcal{P}_{\mathcal{T}(X)}(B_iXA_i)\rangle \\ = \; & \mathbb{E}[\mathrm{tr}(A_j'V\Sigma_R U'B_j'B_iU\Sigma_R V'A_i)] \\ & - \mathbb{E}[\mathrm{tr}(A_j'V\Sigma_R U'B_j'UI_NU'B_iU\Sigma_R V'A_iVI_NV')]. \end{aligned}$$

By cyclicity of trace and iterated expectations we have

$$\begin{aligned} & \mathbb{E}[\mathrm{tr}(A_j' V \Sigma_R U' B_j' U I_N U' B_i U \Sigma_R V' A_i V I_N V')] \\ = \; & \mathbb{E}_U[\mathrm{tr}(\Sigma_R^{1/2} U' B_j' U I_N U' B_i U \Sigma_R^{1/2} (\mathbb{E}_V[\Sigma_R^{1/2} V' A_i V I_N V' A_j' V \Sigma_R^{1/2}]))] = 0, \end{aligned}$$

which proves part (ii) of the proof. It leaves to prove the claim. We do so by verifying that the matrix $\mathbb{E}[V' A_i V E_{11} V' A_j' V]$ is 0 in every coordinate, which is equivalent to showing that $\mathbb{E}(\mathbf{v}_m' A_i \mathbf{v}_1)(\mathbf{v}_n' A_j \mathbf{v}_1) = 0$ for all m, n. There are three cases.

[Case $m \neq n$]: Without loss of generality suppose that $m \neq 1$. Then $\mathbb{E}(\mathbf{v}_m' A_i \mathbf{v}_1)(\mathbf{v}_n' A_j \mathbf{v}_1) = \mathbb{E}[\mathbb{E}[(\mathbf{v}_m' A_i \mathbf{v}_1)(\mathbf{v}_n' A_j \mathbf{v}_1) | \mathbf{v}_1, \mathbf{v}_n]] = 0$.

[Case $m = n = 1$]: We divide into further sub-cases depending on the subspaces A_i, A_j belong to. If $A_i \in \mathcal{H}$ then $\mathbf{v}_1' A_i \mathbf{v}_1 = 0$ since it is a scalar. Hence we eliminate the case where either matrix is in $\mathcal{H}$. Since $\langle A_i, A_j \rangle = 0$ it cannot be that both $A_i, A_j \in \mathcal{I}$. Suppose that $A_i = I/\sqrt{p}$ and $A_j \in \mathcal{G}$. Then $\mathbb{E}[(\mathbf{v}_1' A_i \mathbf{v}_1)(\mathbf{v}_1' A_j \mathbf{v}_1)] = \mathbb{E}[(\mathbf{v}_1' A_j \mathbf{v}_1)]/\sqrt{p} = \mathbb{E}[\mathrm{tr}(A_j \mathbf{v}_1 \mathbf{v}_1')]/\sqrt{p} = 0$. Our remaining case is when $A_i, A_j \in \mathcal{G}$, and $\langle A_i, A_j \rangle = 0$. As before we let $\mathbf{v}_1 = (v_1, \ldots, v_q)'$. Then

$$\begin{aligned} & \mathbb{E}[(\mathbf{v}_1' A_i \mathbf{v}_1)(\mathbf{v}_1' A_j \mathbf{v}_1)] = \mathbb{E}[\sum_{pqrs} A_{i,pq} A_{j,rs} v_p v_q v_r v_s] \\ = \; & \sum_p A_{i,pp} A_{j,pp} \mathbb{E}[v_p^4] + \sum_{p \neq r} A_{i,pp} A_{j,rr} \mathbb{E}[v_p^2 v_r^2] + 2 \sum_{p \neq q} A_{i,pq} A_{j,pq} \mathbb{E}[v_p^2 v_p^2], \end{aligned}$$

where in the second equality we used the fact that A_i, A_j are symmetric to obtain a factor of 2 in the last term. Next we apply the relations $\mathbb{E}[v_p^4] = 3/(q(q+2))$, $\mathbb{E}[v_p^2 v_r^2] = 1/(q(q+2))$, as well as the relations $0 = \langle A_i, I \rangle \langle A_j, I \rangle = \sum_p A_{i,pp} A_{j,pp} + \sum_{p \neq r} A_{i,pp} A_{j,rr}$, and $0 = \langle A_i, A_j \rangle = \sum_p A_{i,pp} A_{j,pp} + \sum_{p \neq q} A_{i,pq} A_{j,pq}$ to conclude that the mean is zero.

[Case $m = n \neq 1$]: We have

$$\begin{aligned} & \mathbb{E}[(\mathbf{v}_m' A_i \mathbf{v}_1)(\mathbf{v}_m' A_j \mathbf{v}_1)] \\ = \; & \mathbb{E}[\mathbb{E}[\mathrm{tr}(A_i \mathbf{v}_1 \mathbf{v}_1' A_j' \mathbf{v}_m \mathbf{v}_m')] | \mathbf{v}_1] \\ = \; & \mathbb{E}[\mathrm{tr}(A_i \mathbf{v}_1 \mathbf{v}_1' A_j' (I - \mathbf{v}_1 \mathbf{v}_1')/(p-1)) | \mathbf{v}_1] \\ = \; & \mathbb{E}[\mathrm{tr}(A_i \mathbf{v}_1 \mathbf{v}_1' A_j' / (p-1))] = \mathbb{E}[\mathrm{tr}(A_i I A_j' / (p(p-1)))] = 0, \end{aligned}$$

where the first equality applies the fact that, conditioned on $\mathbf{v}_1$, $\mathbb{E}[\mathbf{v}_m \mathbf{v}_m']$ is the identity matrix in the subspace $\mathcal{T}(\mathbf{v}_1 \mathbf{v}_1')^\perp$ suitably scaled, and the second inequality applies the previous case. □

Proof of Proposition 2.3.8. First we have $0 \preceq X^{(j)} \boxtimes X^{(j)} \preceq s^2 r\mathsf{I}$. By Proposition A.2.1 we have $\mathbb{E}[X^{(j)} \boxtimes X^{(j)}] = (\bar{s}^2 r/p^2)\mathsf{I}$. Since $(X^{(j)} \boxtimes X^{(j)} - (\bar{s}^2 r/p^2)\mathsf{I})^2 \preceq s^4 r^2 \mathsf{I}$, we have $\mathbb{P}(\|(1/n)\sum_{i=1}^n X^{(j)} \boxtimes X^{(j)} - (\bar{s}^2 r/p^2)\mathsf{I}\| > trs^2) \leq 2p\exp(-t^2 n/8)$ via an application of the Matrix Hoeffding inequality (Theorem 1.3 in [146]).

Second we have $\|X^{(j)} \boxtimes X^{(j)}\|_2 \leq s^2 r$, and $\|\mathcal{P}_{\mathcal{T}(X^{(j)})}\|_2 = 1$, and hence $(X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})} \preceq s^2 r\mathsf{I} \otimes \mathsf{I} =: s^2 r\mathtt{I}$. From Proposition A.2.2 we have

$$\mathbb{E}[(X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})}] \preceq \frac{\bar{s}^2 r}{p^2}\mathtt{I}_{\mathcal{W}} + \frac{16\bar{s}^2 r^2}{p^3}\mathtt{I}_{\mathcal{W}^\perp}.$$

Since $((X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})} - r\mathtt{I})^2 \preceq s^4 r^2 \mathtt{I}$ we have

$$\begin{aligned} &\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n (X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})} - \mathbb{E}[(X^{(j)} \boxtimes X^{(j)}) \otimes \mathcal{P}_{\mathcal{T}(X^{(j)})}]\right) \geq trs^2\right) \\ \leq\ & p\exp(-t^2 n/8) \end{aligned}$$

by an application of the Matrix Hoeffding inequality.

Let $t = t_1/(5p^2)$ in the first concentration bound, and $t = t_2/(5p^2)$ in the second concentration bound. Then $\Delta(\mathscr{X}) \leq t_1 s^2 r/(5p^2)$, and $\Omega(\mathscr{X}) \leq 16s^2 r^2/p^3 + t_2 s^2 r/(5p^2)$, with probability greater than $1 - 2p\exp(-nt_1^2/(200p^4)) - p\exp(-nt_2^2/(200p^4))$. We condition on the event that both inequalities hold. Since $\Delta(\mathscr{X}) \leq t_1 s^2 r/(5p^2) \leq s^2 r/(20p^2)$, by Lemma 2.3.10 we have $\Lambda(\mathscr{X}) \geq s^2 r/(5p^2)$, and hence $\Delta(\mathscr{X})/\Lambda(\mathscr{X}) \leq t_1$, and $\Omega(\mathscr{X})/\Lambda(\mathscr{X}) \leq 80r/p + t_2$. □

A.3 Stability of Matrix and Operator Scaling

In this section we prove a stability property of Sinkhorn scaling and Operator Sinkhorn scaling. For Sinkhorn scaling, we show that if a matrix is close to being doubly stochastic and has entries that are suitably bounded away from 0, then the resulting row and column scalings are close to $\mathbf{1} := (1, \ldots, 1)'$. We also prove the operator analog of this result. These results are subsequently used to prove Propositions 2.3.1 and 2.3.7. We note that there is an extensive literature on the stability of matrix scaling, with results similar to ours. However, Proposition A.3.1 in this section is stated in a manner that is directly suited to our analysis, and we include it for completeness.

A.3.1 Main results

Proposition A.3.1 (Local stability of Matrix Scaling) *Let $T \in \mathbb{R}^{p\times p}$ be a matrix such that*

1. $|\langle \mathbf{e}_i, T(\mathbf{e}_j)\rangle - 1/p| \le 1/(2p)$ *for all standard basis vectors* $\mathbf{e}_i, \mathbf{e}_j$*; and*

2. $\epsilon := \max\{\|T\mathbf{1} - \mathbf{1}\|_\infty, \|T'\mathbf{1} - \mathbf{1}\|_\infty\} \le 1/(48\sqrt{p})$.

Let D_1, D_2 be diagonal matrices such that $D_2 T D_1$ is doubly stochastic. Then

$$\|D_2 \otimes D_1 - \mathsf{I}\|_2 \le 96\sqrt{p}\epsilon.$$

Proposition A.3.2 (Local stability of Operator Scaling) *Let $\mathsf{T} : \mathbb{S}^p \to \mathbb{S}^p$ be a rank-indecomposable linear operator such that*

1. $|\langle \mathbf{v}\mathbf{v}', \mathsf{T}(\mathbf{u}\mathbf{u}')\rangle - 1/p| \le 1/(2p)$ *for all unit-norm vectors* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$*; and*

2. $\epsilon := \max\{\|\mathsf{T}(I) - I\|_2, \|\mathsf{T}'(I) - I\|_2\} \le 1/(48\sqrt{p})$.

Let $N_1, N_2 \in \mathbb{S}^p$ be positive definite matrices such that $(N_2 \otimes N_2) \circ \mathsf{T} \circ (N_1 \otimes N_1)$ is doubly stochastic. Then $\|N_2^2 \otimes N_1^2 - \mathsf{I}\|_2 \le 96\sqrt{p}\epsilon$. Furthermore we have $\|N_2 \otimes N_1 - \mathsf{I}\|_2 \le 96\sqrt{p}\epsilon$.

A.3.2 Proofs

The proof of Proposition A.3.1 relies on the fact that matrix scaling can be cast as the solution of a convex program; specifically, we utilize the correspondence between diagonal matrices D_1, D_2 such that $D_2 T D_1$ is doubly stochastic, and the vectors $\boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_p)', \boldsymbol{\eta} := (\eta_1, \ldots, \eta_p)'$ that minimize the following convex function

$$F(\boldsymbol{\varepsilon}, \boldsymbol{\eta}) = \sum_{ij} T_{ij} \exp(\varepsilon_i + \eta_j) - \sum_i \varepsilon_i - \sum \eta_j$$

via the maps $(D_2)_{ii} = \exp(\varepsilon_i)$ and $(D_1)_{jj} = \exp(\eta_j)$ [67] (see also [85]) – this holds for all matrices T with positive entries. We remark that one can derive the above relationship from first order optimality. In the following we prove bounds on the minima of F (see Lemma A.3.6).

The proof of Proposition A.3.2 relies on a reduction to the set-up in Proposition A.3.1.

We begin with a lower estimate of the sum of exponential functions. We use the estimate to prove Proposition A.3.1.

Definition A.3.3 *Let $\alpha \geq 0$. Define the function $c_\alpha : \mathbb{R} \to \mathbb{R}$*

$$c_\alpha(x) = \begin{cases} \frac{1}{2}\exp(-\alpha)x^2 & \text{if } |x| \leq \alpha \\ \frac{1}{2}\exp(-\alpha)\alpha|x| & \text{if } |x| \geq \alpha \end{cases}$$

Remark. *Note that the function $c_\alpha(\cdot)$ is continuous.*

Lemma A.3.4 *For all x*

$$\exp(x) \geq 1 + x + c_\alpha(x).$$

Proof of Lemma A.3.4. The second derivative of $\exp(x)$ is $\exp(x)$, and it is greater than $\exp(-\alpha)$ over all x such that $|x| \leq \alpha$. Hence, by strong convexity of $\exp(x)$, we have $\exp(x) \geq 1 + x + (1/2)\exp(-\alpha)x^2$ over the interval $[-\alpha, \alpha]$.

It follows that $\exp(\alpha) \geq 1 + \alpha + c_\alpha(\alpha)$, and $\exp(-\alpha) \geq 1 - \alpha + c_\alpha(-\alpha)$. Since the function $\exp(x)$ is convex, and c_α is linear in the intervals $(-\infty, -\alpha]$ and $[\alpha, \infty)$ respectively, it suffices to check that (i) the gradient of $\exp(x)$ at $x = \alpha$, which is $\exp(\alpha)$, exceeds that of $c_\alpha(\cdot)$, and (ii) the gradient of $c_\alpha(\cdot)$ exceeds that of $\exp(x)$ at $x = -\alpha$, which is $\exp(-\alpha)$.

First we prove (i). Since $\alpha \geq 0$ we have $1 + 2\alpha \geq \sqrt{1 + 2\alpha}$. Hence $2\exp(\alpha) \geq 2 + 2\alpha \geq 1 + \sqrt{1 + 2\alpha}$. By noting that the quadratic $2z^2 - 2z - \alpha = 0$ has roots $(1/2) \pm (1/2)\sqrt{1 + 2\alpha}$, we have the inequality $\exp(\alpha) \geq 1 + (1/2)\exp(-\alpha)\alpha$, from which (i) follows.

Next we prove (ii). Since $\alpha \geq 0$, we have $\exp(\alpha) \geq 1 + \alpha \geq 1 + \alpha/2$, and hence $1 - (1/2)\exp(-\alpha)\alpha \geq \exp(-\alpha)$ from which (ii) follows. □

Lemma A.3.5 *Let $\{\varepsilon_i\}_{i=1}^p$ and $\{\eta_j\}_{j=1}^p$ be a collection of reals satisfying $(\sum_i \varepsilon_i) + (\sum_j \eta_j) \geq -2p$. Then there is a constant $d \in \mathbb{R}$ for which*

$$\frac{1}{p}\sum_{ij} \exp(\varepsilon_i + \eta_j) \geq p + \left(\sum_i (\varepsilon_i + c_\alpha(\varepsilon_i + d))\right) + \left(\sum_j (\eta_j + c_\alpha(\eta_j - d))\right).$$

Proof. Consider the function

$$f(d) := \sum_i (\varepsilon_i + d + c_\alpha(\varepsilon_i + d)) - \sum_j \left(\eta_j - d + c_\alpha(\eta_j - d)\right).$$

Then $f(\cdot)$ is continuous in d, and $f(d) \to \pm\infty$ as $d \to \pm\infty$. By the Intermediate Value Theorem, there is a $d^\star$ for which $f(d^\star) = 0$. Then

$$\sum_i \left(1 + \varepsilon_i + d^\star + c_\alpha(\varepsilon_i + d^\star)\right) = \sum_j \left(1 + \eta_j - d^\star + c_\alpha(\eta_j - d^\star)\right).$$

By summing both sides and noting that $c_\alpha(\cdot) \geq 0$, we have that each side of the above equation is nonnegative. It follows that

$$\begin{aligned}
& \frac{1}{p}\sum_{ij} \exp(\varepsilon_i + \eta_j) \\
= \;& \frac{1}{p}\left(\sum_i \exp(\varepsilon_i + d^\star)\right)\left(\sum_j \exp(\eta_j - d^\star)\right) \\
\geq \;& \frac{1}{p}\left(\sum_i \left(1 + \varepsilon_i + d^\star + c_\alpha(\varepsilon_i + d^\star)\right)\right)\left(\sum_j \left(1 + \eta_j - d^\star + c_\alpha(\eta_j - d^\star)\right)\right) \\
\geq \;& p + \left(\sum_i (\varepsilon_i + c_\alpha(\varepsilon_i + d))\right) + \left(\sum_j (\eta_j + c_\alpha(\eta_j - d))\right).
\end{aligned}$$

□

Lemma A.3.6 *Given vectors $\boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_p)$ and $\boldsymbol{\eta} := (\eta_1, \ldots, \eta_p)$ define*

$$F(\boldsymbol{\varepsilon}, \boldsymbol{\eta}) = \sum_{ij} T_{ij} \exp(\varepsilon_i + \eta_j) - \sum_i \varepsilon_i - \sum_j \eta_j, \tag{A.2}$$

and $\epsilon_{ij} := T_{ij} - 1/p$. Suppose (i) $|\epsilon_{ij}| \leq 1/2p$, and (ii) $\epsilon := \max\{|\sum_i \epsilon_{ij}|, |\sum_j \epsilon_{ij}|\} \leq 1/(24\sqrt{p})$. Let $\boldsymbol{\varepsilon}^\star, \boldsymbol{\eta}^\star$ be a minimizer of F. Then $|\varepsilon_i^\star + \eta_j^\star| \leq 48\sqrt{p}\epsilon$, for all i, j.

Proof. Suppose $|\varepsilon_i + \eta_j| > 48\sqrt{p}\epsilon$ for some (i, j). We show that $\boldsymbol{\varepsilon}, \boldsymbol{\eta}$ cannot be a minimum. We split the analysis to two cases.

$[(\sum_i \varepsilon_i) + (\sum_j \eta_j) < -2p]$: Since $T_{ij} > 0$ we have $F(\boldsymbol{\varepsilon}, \boldsymbol{\eta}) > -(\sum_i \varepsilon_i) - (\sum_j \eta_j) \geq 2p$. Then $F(\mathbf{0}, \mathbf{0}) = \sum_i (\sum_j T_{ij}) = \sum_i (1 + \sum_j \epsilon_{ij}) \leq p(1 + 1/(24\sqrt{p})) \leq 2p < F(\boldsymbol{\varepsilon}, \boldsymbol{\eta})$.

$[(\sum_i \varepsilon_i) + (\sum_j \eta_j) \geq -2p]$: Let $\alpha = 24\sqrt{p}\epsilon$, and define the sets

1. $\mathscr{S}(\boldsymbol{\varepsilon}) = \{i : |\varepsilon_i| \geq \alpha\}$;
2. $\mathscr{T}(\boldsymbol{\varepsilon}) = \{i : \alpha > |\varepsilon_i| \geq 4\epsilon \exp(\alpha)\}$; and
3. $\mathscr{U}(\boldsymbol{\varepsilon}) = \{i : 4\epsilon \exp(\alpha) > |\varepsilon_i|\}$.

Similarly define the sets $\mathscr{S}(\boldsymbol{\eta}), \mathscr{T}(\boldsymbol{\eta}), \mathscr{U}(\boldsymbol{\eta})$.

First since $\alpha \leq 1$, we have $\alpha \geq \alpha \exp(\alpha)/3 \geq 8\sqrt{p}\epsilon \exp(\alpha) \geq 8\epsilon \exp(\alpha)$, and hence

$$\frac{1}{4}\left(\sum_{i\in\mathscr{S}(\boldsymbol{\varepsilon})} c_\alpha(\varepsilon_i) + \sum_{j\in\mathscr{S}(\boldsymbol{\eta})} c_\alpha(\eta_j)\right) \geq \epsilon\left(\sum_{i\in\mathscr{S}(\boldsymbol{\varepsilon})} |\varepsilon_i| + \sum_{j\in\mathscr{S}(\boldsymbol{\eta})} |\eta_j|\right).$$

Second

$$\begin{aligned}\frac{1}{2}\left(\sum_{i\in\mathscr{T}(\boldsymbol{\varepsilon})} c_\alpha(\varepsilon_i) + \sum_{j\in\mathscr{T}(\boldsymbol{\eta})} c_\alpha(\eta_j)\right) &= \sum_{i\in\mathscr{T}(\boldsymbol{\varepsilon})} \frac{1}{4}\exp(-\alpha)\varepsilon_i^2 + \sum_{j\in\mathscr{T}(\boldsymbol{\eta})} \frac{1}{4}\exp(-\alpha)\eta_j^2 \\ &\geq \epsilon\left(\sum_{i\in\mathscr{T}(\boldsymbol{\varepsilon})} |\varepsilon_i| + \sum_{j\in\mathscr{T}(\boldsymbol{\eta})} |\eta_j|\right).\end{aligned}$$

Third since there is an index (i, j) such that $|\varepsilon_i + \eta_j| > 48\sqrt{p}\epsilon$, one of the sets $\mathscr{S}(\boldsymbol{\varepsilon}), \mathscr{S}(\boldsymbol{\eta})$ is nonempty. By noting that $\alpha \exp(-\alpha) \geq 8\sqrt{p}\epsilon$, we have

$$\frac{1}{4}\left(\sum_{i\in\mathscr{S}(\boldsymbol{\varepsilon})} c_\alpha(\varepsilon_i) + \sum_{j\in\mathscr{S}(\boldsymbol{\eta})} c_\alpha(\eta_j)\right) > \epsilon \times 2p \times 4\epsilon \exp(\alpha) \geq \epsilon\left(\sum_{i\in\mathscr{U}(\boldsymbol{\varepsilon})} |\varepsilon_i| + \sum_{j\in\mathscr{U}(\boldsymbol{\eta})} |\eta_j|\right).$$

We have $\epsilon(\sum_i |\varepsilon_i| + \sum_j |\eta_j|) \geq \sum_i(\varepsilon_i(\sum_j \epsilon_{ij})) + \sum_j(\eta_j(\sum_i \epsilon_{ij})) = \sum_{ij} \epsilon_{ij}(\varepsilon_i + \eta_j)$. By combining the above inequalities with Lemma A.3.5 we have

$$\frac{1}{2p}\sum\left(\exp(\varepsilon_i + \eta_j) - (\varepsilon_i + \eta_j) - 1\right) \geq \frac{1}{2}\left(\sum_i c_\alpha(\varepsilon_i) + \sum_j c_\alpha(\eta_j)\right) > \sum_{ij} \epsilon_{ij}(\varepsilon_i + \eta_j). \tag{A.3}$$

Also, since $\exp(\varepsilon_i + \eta_j) - (\varepsilon_i + \eta_j) - 1 \geq 0$ for all i, j, and $|\epsilon_{ij}| \leq 1/(2p)$, we have

$$\begin{aligned}\frac{1}{2p}\sum_{ij}(\exp(\varepsilon_i + \eta_j) - (\varepsilon_i + \eta_j) - 1) &\geq \max_{ij}|\epsilon_{ij}| \times \sum_{ij}\left|\exp(\varepsilon_i + \eta_j) - (\varepsilon_i + \eta_j) - 1\right| \\ &\geq \sum_{ij} \epsilon_{ij}(\exp(\varepsilon_i + \eta_j) - (\varepsilon_i + \eta_j) - 1). \end{aligned} \tag{A.4}$$

By combining equations (A.3) and (A.4) we have

$$\frac{1}{p}\sum_{ij}(\exp(\varepsilon_i + \eta_j) - (\varepsilon_i + \eta_j) - 1) > -\sum_{ij} \epsilon_{ij}(\exp(\varepsilon_i + \eta_j) - 1),$$

which implies $F(\boldsymbol{\varepsilon}, \boldsymbol{\eta}) > F(\mathbf{0}, \mathbf{0})$. □

Proof of Proposition A.3.1. By Lemma A.3.6 any minimum $\boldsymbol{\varepsilon}^\star, \boldsymbol{\eta}^\star$ satisfies $|\varepsilon_i^\star + \eta_j^\star| \leq 48\sqrt{p}\epsilon$. Hence by the one-to-one correspondence between the minima of F and the diagonal scalings D_1, D_2 [67], we have $\|D_2 \otimes D_1 - \mathrm{I}\|_2 \leq \exp(48\sqrt{p}\epsilon) - 1 \leq 96\sqrt{p}\epsilon$. □

Proof of Proposition A.3.2. Without loss of generality we may assume that N_1, N_2 are diagonal matrices, say D_1, D_2 respectively. Define the matrix $T_{ij} = \langle \mathbf{e}_i \mathbf{e}_i', \mathsf{T}(\mathbf{e}_j \mathbf{e}_j') \rangle$. It is straightforward to check that T satisfies the conditions of Proposition A.3.1; moreover, the condition that $(N_2 \otimes N_2) \circ \mathsf{T} \circ (N_1 \otimes N_1)$ is a doubly stochastic operator implies that $D_2^2 T D_1^2$ is a doubly stochastic matrix. By Proposition A.3.1 we have $\|D_1^2 \otimes D_2^2 - \mathsf{I}\|_2 \leq 96\sqrt{p}\epsilon$, and hence $\|N_1^2 \otimes N_2^2 - \mathsf{I}\|_2 \leq 96\sqrt{p}\epsilon$. Since N_1, N_2 are self-adjoint, we also have $\|N_1 \otimes N_2 - \mathsf{I}\|_2 \leq 96\sqrt{p}\epsilon$. □

A.4 Proof of Proposition 2.3.7

In this section we prove that Gaussian linear maps that are subsequently normalized satisfy the deterministic conditions in Theorem 2.3.5 concerning the linear map $\mathcal{L}^\star$ with high probability. There are two steps to our proof. First, we state sufficient conditions for linear maps such that, when normalized, satisfy the deterministic conditions. Second, we show that Gaussian maps satisfy these sufficient conditions with high probability.

We introduce the following parameter that measures how close a linear map $\mathcal{L}$ is to being normalized.

Definition A.4.1 *Let $\mathcal{L} \in \mathbb{R}^{p \times p} \to \mathbb{R}^d$ be a linear map. The* nearly normalized parameter *of $\mathcal{L}$ is defined as*

$$\epsilon(\mathcal{L}) := \max\{\|\mathsf{T}_{\mathcal{L}}(I) - I\|_2, \|\mathsf{T}'_{\mathcal{L}}(I) - I\|_2\}.$$

Proposition A.4.2 *Let $\mathcal{L} : \mathbb{R}^{p \times p} \to \mathbb{R}^d$ be a linear map that satisfies (i) the restricted isometry condition $\delta_r(\mathcal{L}) \leq 1/2$, and (ii) whose nearly normalized parameter satisfies $\epsilon(\mathcal{L}) \leq 1/(650\sqrt{p})$. Let $\mathcal{L} \circ \mathsf{N}_{\mathcal{L}}$ be the normalized linear map where $\mathsf{N}_{\mathcal{L}}$ is a positive definite rank-preserver. Then $\mathcal{L} \circ \mathsf{N}_{\mathcal{L}}$ satisfies the restricted isometry condition $\delta_r(\mathcal{L} \circ \mathsf{N}) \leq \bar{\delta_r} := (1 + \delta_r(\mathcal{L}))(1 + 96\sqrt{p}\epsilon(\mathcal{L}))^2 - 1 < 1$. Moreover, $\|\mathcal{L} \circ \mathsf{N}_{\mathcal{L}}\|_2 \leq (1 + 96\sqrt{p}\epsilon(\mathcal{L}))\|\mathcal{L}\|_2$.*

Proof of Proposition A.4.2. Since $\mathcal{L}$ satisfies the restricted isometry condition $\delta_1(\mathcal{L}) \leq 1/2$, we have $|\langle \mathbf{v}\mathbf{v}', \mathsf{T}_{\mathcal{L}}(\mathbf{u}\mathbf{u}') \rangle - 1/p| \leq 1/(2p)$ for all unit-norm vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$. In addition, the linear map $\mathcal{L}$ has nearly normalized parameter $\epsilon(\mathcal{L}) \leq 1/(650\sqrt{p})$. Hence by applying Proposition A.3.2 to the linear map $\mathsf{T}_{\mathcal{L}}$, any pair of positive definite matrices Q_2, Q_1 such that $Q_2 \otimes Q_2 \circ \mathsf{T}_{\mathcal{L}} \circ Q_1 \otimes Q_1$ is doubly stochastic satisfies $\|Q_2 \otimes Q_1 - \mathsf{I}\|_2 \leq 96\sqrt{p}\epsilon(\mathcal{L})$. By noting the correspondence between such matrices

with the positive definite rank-preserver $\mathsf{N}_{\mathcal{L}}$ such that $\mathcal{L} \circ \mathsf{N}_{\mathcal{L}}$ is normalized via the relation $\mathsf{N}_{\mathcal{L}} = Q_2 \otimes Q_1$ (see Corollary 2.2.5), we have $\|\mathsf{N}_{\mathcal{L}}\|_2 \leq 1 + 96\sqrt{p}\epsilon(\mathcal{L})$.

Let X be a matrix with rank at most r. Then

$$\|\mathcal{L}(\mathsf{N}_{\mathcal{L}}(X))\|_{\ell_2} \leq \sqrt{1+\delta_r(\mathcal{L})}\|\mathsf{N}_{\mathcal{L}}\|_2\|X\|_{\ell_2} \leq \sqrt{1+\delta_r(\mathcal{L})}(1+96\sqrt{p}\epsilon(\mathcal{L}))\|X\|_{\ell_2},$$

and hence $\|\mathcal{L}(\mathsf{N}_{\mathcal{L}}(X))\|_{\ell_2}^2 \leq (1+\bar{\delta}_r)\|X\|_{\ell_2}^2$. A similar set of steps show that $\|\mathcal{L}(\mathsf{N}_{\mathcal{L}}(X))\|_{\ell_2}^2 \geq (1-\bar{\delta}_r)\|X\|_{\ell_2}^2$. Last $\|\mathcal{L} \circ \mathsf{N}_{\mathcal{L}}\|_2 \leq \|\mathcal{L}\|_2\|\mathsf{N}_{\mathcal{L}}\|_2 \leq (1+96\sqrt{p}\epsilon)\|\mathcal{L}\|_2$. □

Proposition A.4.3 *([43, Theorem II.13]) Let $t > 0$ be fixed. Suppose $\mathcal{L} \sim \mathcal{N}(0, 1/d)$. Then with probability greater than $1 - \exp(-t^2 d/2)$ we have $\|\mathcal{L}\|_2 \leq \sqrt{p^2/d} + 1 + t$.*

Proposition A.4.4 *([28, Theorem 2.3]) Let $0 < \delta < 1$ be fixed. There exists constants c_1, c_2 such that for $d \geq c_1 pr$, if $\mathcal{L} \sim \mathcal{N}(0, 1/d)$, then with probability greater than $1 - 2\exp(-c_2 d)$ the linear map $\mathcal{L}$ satisfies the restricted isometry condition $\delta_r(\mathcal{L}) \leq \delta$.*

Proposition A.4.5 (Gaussian linear maps are nearly normalized) *Suppose $3/\sqrt{d} \leq \epsilon \leq 3$. Suppose $\mathcal{L} \sim \mathcal{N}(0, 1/d)$. Then with probability greater than $1-4\exp(-p(-1+\sqrt{d}\epsilon/3)^2/2)$ the nearly normalized parameter of $\mathcal{L}$ is smaller than ϵ.*

Bounding the nearly normalized parameter of a Gaussian linear map exactly corresponds to computing the deviation of the sum of independent Wishart matrices from its mean in spectral norm. To do so we appeal to the following concentration bound.

Proposition A.4.6 (Concentration of sum of Wishart Matrices) *Suppose $3/\sqrt{d} \leq t \leq 3$. Let $\{X^{(j)}\}_{j=1}^d, X^{(j)} = G^{(j)}G^{(j)\prime}$, where $G^{(j)} \in \mathbb{R}^{p\times p}, G^{(j)} \sim \mathcal{N}(0, 1/p)$, be a collection of independent Wishart matrices. Then $\mathbb{P}(\|\frac{1}{d}\sum_{j=1}^d X^{(j)} - I\|_2 \geq t) \leq 2\exp(-p(-1+\sqrt{d}t/3)^2/2)$.*

Proof of Proposition A.4.6. Consider the linear map $G = [G^{(1)}|\ldots|G^{(d)}]$. Then $\sum_{j=1}^d X^{(j)} = GG'$, and $\|\frac{1}{d}\sum_{j=1}^d X^{(j)} - I\|_2 \leq t$ if and only if $\sigma(G) \in [\sqrt{d(1-t)}, \sqrt{d(1+t)}]$. By [43, Theorem II.13] we have $\sigma(G) \in [\sqrt{d}-1-\tilde{t}, \sqrt{t}+1+\tilde{t}]$ with probability greater than $1 - 2\exp(-p\tilde{t}^2/2)$. The result follows with the choice of $\tilde{t} = -1 + \sqrt{d}t/3$. □

Proof of Proposition A.4.5. This is a direct application of Proposition A.4.6 with $G^{(j)} = \sqrt{p/d}\mathcal{L}^{(j)}$ and $G^{(j)\prime} = \sqrt{p/d}\mathcal{L}^{(j)}$, followed by a union bound. □

Proof of Proposition 2.3.7. We choose $t = 1/50$ in Proposition A.4.3, $\delta = \delta_{4r}/2$ in Proposition A.4.4, and $\epsilon = \delta/(960\sqrt{p})$ in Proposition A.4.5. Then there are constants c_1, c_2, c_3 such that if $d \geq c_1 rq$, then (i) $\|\tilde{\mathcal{L}}\|_2 \leq \sqrt{p^2/d} + 51/50 \leq (101/50)\sqrt{p^2/d}$, (ii) $\tilde{\mathcal{L}}$ satisfies the restricted isometry condition $\delta_{4r}(\tilde{\mathcal{L}}) \leq \delta_{4r}/2$, and (iii) $\tilde{\mathcal{L}}$ is nearly normalized with parameter $\epsilon(\tilde{\mathcal{L}}) \leq \delta_{4r}/960\sqrt{p}$, with probability greater than $1 - c_2 \exp(-c_3 d)$.

By applying Proposition A.4.2 we conclude that the linear map $\mathcal{L}$ satisfies the restricted isometry condition $\delta_{4r}(\mathcal{L}) \leq (1 + \delta_{4r}/2)(1 + \delta_{4r}/10)^2 - 1 \leq \delta_{4r}$, and $\|\mathcal{L}\|_2 \leq \sqrt{5p^2/d}$. □

A.5 Proof of Proposition 2.3.1

Proof of Proposition 2.3.1. First we check that the linear map $\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E})$ satisfies the restricted isometry condition $\delta_1(\mathcal{L}^\star \circ (\mathsf{I}+\mathsf{E})) \leq 1/2$. For any rank-one unit-norm matrix X we have $\|[\mathcal{L}^\star \circ (\mathsf{I}+\mathsf{E})](X)\|_{\ell_2} \leq \|\mathcal{L}^\star(X)\|_{\ell_2} + \|\mathcal{L}^\star(\mathsf{E}(X))\|_{\ell_2} \leq \sqrt{1 + 1/10} + 1/150 \leq \sqrt{1 + 1/2}$. A similar set of inequalities show that $\|[\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E})](X)\|_{\ell_2} \geq \sqrt{1 - 1/2}$.

Second we check that the nearly normalized parameter of $\mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E})$ satisfies $\epsilon(\mathcal{L}^\star \circ (\mathsf{I}+\mathsf{E})) \leq 1/48\sqrt{p}$. Denote $\mathcal{E} := \mathcal{L}^\star \circ \mathsf{E}$. For all unit-norm rank-one matrices E we have $\|\mathcal{E}(E)\|_2^2 \leq \|\mathcal{L}^\star\|_2^2 \|\mathsf{E}\|_{\ell_2}^2$. Hence for any unit-norm $\mathbf{u} \in \mathbb{R}^p$ we have

$$\frac{1}{p}\sum_{j=1}^{d} \langle \mathcal{E}_j \mathcal{E}_j', \mathbf{u}\mathbf{u}' \rangle = \frac{1}{p}\sum_{j=1}^{d}\sum_{k=1}^{p} (\mathcal{E}_j' \mathbf{u})_k^2 = \frac{1}{p}\sum_{k=1}^{p} \|\mathcal{E}(\mathbf{u}\mathbf{e}_k')\|_{\ell_2}^2 \leq \|\mathcal{L}^\star\|_2^2 \|\mathsf{E}\|_{\ell_2}^2.$$

Using the fact that $\mathcal{L}^\star$ is normalized we have

$$\frac{1}{p}\sum_{j=1}^{d} \langle \mathcal{L}_j^\star \mathcal{L}_j^{\star\prime}, \mathbf{u}\mathbf{u}' \rangle = 1.$$

By combining the previous inequalities with an application of Cauchy-Schwarz we have

$$\begin{aligned}
& \langle \mathsf{T}_{\mathcal{L}^\star \circ (\mathsf{I}+\mathsf{E})}(I) - I, \mathbf{u}\mathbf{u}' \rangle \\
= \; & \langle \mathsf{T}_{\mathcal{L}^\star + \mathcal{E}}(I) - \mathsf{T}_{\mathcal{L}^\star}(I), \mathbf{u}\mathbf{u}' \rangle \\
= \; & \frac{1}{p}\sum_{j=1}^{d} \langle \mathcal{E}_j \mathcal{E}_j', \mathbf{u}\mathbf{u}' \rangle + \frac{1}{p}\sum_{j=1}^{d} \langle \mathcal{L}_j^\star \mathcal{E}_j', \mathbf{u}\mathbf{u}' \rangle + \frac{1}{p}\sum_{j=1}^{d} \langle \mathcal{E}_j \mathcal{L}_j^{\star\prime}, \mathbf{u}\mathbf{u}' \rangle \\
\leq \; & 3\|\mathcal{L}^\star\|_2 \|\mathsf{E}\|_{\ell_2},
\end{aligned}$$

Further more since $\mathbf{u}$ is arbitrary it follows that

$$\|\mathsf{T}_{\mathcal{L}^\star+\mathcal{E}}(I) - I\|_2 \leq 3\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2}.$$

Using a similar sequence of steps one can show that $\|\mathsf{T}'_{\mathcal{L}^\star+\mathcal{E}}(I)-I\|_2 \leq 3\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2}$. Thus $\epsilon(\mathcal{L}^\star \circ (\mathsf{I}+\mathsf{E})) \leq 3\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2} \leq 1/(48\sqrt{p})$.

The result follows by applying Proposition A.3.2 to the linear map $\mathsf{T}_{\mathcal{L}^\star\circ(\mathsf{I}+\mathsf{E})}$. □

A.6 Proof of Proposition 2.3.2

The proof of Proposition 2.3.2 is based on the following result concerning affine rank minimization, which may be of independent interest.

Proposition A.6.1 *Suppose $X^\star$ is a $p \times p$ rank-r matrix satisfying $\sigma_r(X^\star) \geq 1/2$. Let $\mathbf{y} = \mathcal{L}(X^\star)+\mathbf{z}$, where the linear map $\mathcal{L}$ satisfies the restricted isometry condition $\delta_{4r}(\mathcal{L}) \leq 1/10$, and $\|\mathcal{L}'\mathbf{z}\|_2 =: \epsilon \leq 1/(80r^{3/2})$. Let $\hat{X}$ be the optimal solution to*

$$\hat{X} \;=\; \underset{X}{\operatorname{argmin}} \|\mathbf{y} - \mathcal{L}(X)\|_{\ell_2}^2 \qquad \text{s.t.} \qquad \operatorname{rank}(X) \leq r.$$

Then (i) $\|\hat{X}-X^\star\|_2 \leq 4\sqrt{r}\epsilon$, and (ii) $\hat{X}-X^\star = [(\mathcal{L}'_{\mathcal{T}(X^\star)}\mathcal{L}_{\mathcal{T}(X^\star)})^{-1}]_{\mathbb{R}^{p\times p}}(\mathcal{L}'_{\mathcal{T}(X^\star)}\mathbf{z})+G$, where $\|G\|_{\ell_2} \leq 340r^{3/2}\epsilon^2$.

The proof of Proposition A.6.1 requires two preliminary results which we state and prove first. Our development relies on results from matrix perturbation theory; we refer the reader to [84, 137] for detailed expositions. Several of our results are minor modifications of analogous results in [34].

The following result and the accompanying proof are minor modifications of Proposition 2.2 in the supplementary material (s.m.) of [34] and its proof. The modification allows us to provide a bound that does not scale with the ambient dimension.

Proposition A.6.2 *Let $X_1, X_2 \in \mathbb{R}^{p\times p}$ be rank-r matrices. Let σ be the smallest nonzero singular value of X_1, and suppose that $\|X_1 - X_2\|_2 \leq \sigma/8$. Then $\|\mathcal{P}_{\mathcal{T}(X_1)^\perp}(X_2)\|_{\ell_2} \leq \sqrt{r}\|X_1 - X_2\|_2^2/(3\sigma)$, and $\|\mathcal{P}_{\mathcal{T}(X_1)^\perp}(X_2)\|_2 \leq \|X_1 - X_2\|_2^2/(5\sigma)$.*

In the following proof, given a matrix $X \in \mathbb{R}^{p\times p}$, we denote $\tilde{X} := \begin{pmatrix} 0 & X' \\ X & 0 \end{pmatrix}$.

Proof of Proposition A.6.2. Let $\tilde{\Delta} = \tilde{X}_2 - \tilde{X}_1$, and let $\kappa = \sigma/4$. By combining equation (1.5) in the s.m. of [34] with the proofs of Propositions 1.2 and 2.2 in the s.m. of [34] it can be shown that $\mathcal{P}_{\mathcal{T}(\tilde{X}_1)^\perp}(\tilde{X}_2) = (1/(2\pi i)) \oint_{C_\kappa} \zeta[\tilde{X}_1 - \zeta I]^{-1}\tilde{\Delta}[\tilde{X}_1 - \zeta I]^{-1}\tilde{\Delta}[\tilde{X}_2 - \zeta I]^{-1} d\zeta$, where the contour integral is taken along C_κ defined as the circle centered at the origin with radius κ.

By a careful use of the inequality $\|AB\|_{\ell_2} \leq \|A\|_2\|B\|_{\ell_2}$, we have $\|[\tilde{X}_1 - \zeta I]^{-1}\tilde{\Delta}[\tilde{X}_1 - \zeta I]^{-1}\tilde{\Delta}[\tilde{X}_2 - \zeta I]^{-1}\|_{\ell_2} \leq \|[\tilde{X}_1 - \zeta I]^{-1}\|_2\|\tilde{\Delta}\|_{\ell_2}\|[\tilde{X}_1 - \zeta I]^{-1}\|_2\|\tilde{\Delta}\|_2\|[\tilde{X}_2 - \zeta I]^{-1}\|_2$. Since $\tilde{\Delta}$ is a matrix with rank at most $4r$, we have $\|\tilde{\Delta}\|_{\ell_2} \leq \sqrt{4r}\|\tilde{\Delta}\|_2$. We proceed to apply the same bounds as those used in the proof of Proposition 1.2 in the s.m. of [34] to obtain $\|\mathcal{P}_{\mathcal{T}(\tilde{X}_1)^\perp}(\tilde{X}_2)\|_{\ell_2} \leq 2\sqrt{r}\kappa^2\|\tilde{\Delta}\|_2^2/((\sigma-\kappa)^2(\sigma - 3\kappa/2)) \leq \sqrt{2r}\|\tilde{X}_1 - \tilde{X}_2\|_2^2/(3\sigma)$. The first inequality follows by noting that $\sqrt{2}\|\mathcal{P}_{\mathcal{T}(X_1)^\perp}(X_2)\|_{\ell_2} = \|\mathcal{P}_{\mathcal{T}(\tilde{X}_1)^\perp}(\tilde{X}_2)\|_{\ell_2}$, and that $\|X_1 - X_2\|_2 = \|\tilde{X}_1 - \tilde{X}_2\|_2$.

The proof of the second inequality follows from a similar argument. □

We define the following distance measure between two subspaces $\mathcal{T}_1$ and $\mathcal{T}_2$ [34]

$$\rho(\mathcal{T}_1, \mathcal{T}_2) := \sup_{\|N\|_2 \leq 1} \|[\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2}](N)\|_2.$$

This definition is useful for quantifying the distance between tangent spaces with respect to the variety of low-rank matrices for pairs of nearby matrices.

Lemma A.6.3 *Let $X_1, X_2 \in \mathbb{R}^{p\times p}$ be matrices with rank at most r, and satisfy $\|X_1 - X_2\|_2 \leq \sigma/8$, where σ is the smallest nonzero singular value value of X_2. Let $\mathcal{T}_1 := \mathcal{T}(X_1)$ and $\mathcal{T}_2 := \mathcal{T}(X_2)$ be tangent spaces on the variety of matrices with rank at most r at the points X_1 and X_2 respectively. Let $\mathcal{L}$ be a linear map satisfying the restricted isometry condition $\delta_{4r}(\mathcal{L}) \leq 1/10$. If $Z_i \in \mathcal{T}_i$, $i \in \{1,2\}$, then $\|[(\mathcal{L}'_{\mathcal{T}_1}\mathcal{L}_{\mathcal{T}_1})^{-1}]_{\mathbb{R}^{p\times p}}(Z_1) - [(\mathcal{L}'_{\mathcal{T}_2}\mathcal{L}_{\mathcal{T}_2})^{-1}]_{\mathbb{R}^{p\times p}}(Z_2)\|_{\ell_2} \leq (43/10)\sqrt{r}\|Z_1 - Z_2\|_2 + 16r\|X_1 - X_2\|_2\|Z_2\|_2/\sigma$.*

Proof of Lemma A.6.3. To simplify notation we denote $Y_i = [(\mathcal{L}'_{\mathcal{T}_i}\mathcal{L}_{\mathcal{T}_i})^{-1}]_{\mathbb{R}^{p\times p}}(Z_i)$, $i \in \{1,2\}$. From the triangle inequality we have $\|Y_1 - Y_2\|_{\ell_2} \leq \|\mathcal{P}_{\mathcal{T}_1^\perp}(Y_1 - Y_2)\|_{\ell_2} + \|\mathcal{P}_{\mathcal{T}_1}(Y_1 - Y_2)\|_{\ell_2}$. We bound both components separately.

$[\|\mathcal{P}_{\mathcal{T}_1^\perp}(Y_1 - Y_2)\|_{\ell_2}]$: From Proposition 2.1 of the s.m. of [34] we have $\rho(\mathcal{T}_1, \mathcal{T}_2) \leq \frac{2}{\sigma}\|X_1 - X_2\|_2$. From Lemma 2.3.9 we have $\|Y_2 - Z_2\|_{\ell_2} \leq \delta_{4r}\|Y_2\|_{\ell_2} \leq \frac{\delta_{4r}}{1-\delta_{4r}}\|Z_2\|_{\ell_2} \leq$

$\frac{\sqrt{2r}\delta_{4r}}{1-\delta_{4r}}\|Z_2\|_2$. Hence

$$\begin{aligned}\|\mathcal{P}_{\mathcal{T}_1^\perp}(Y_2 - Z_2)\|_{\ell_2} &= \|[\mathsf{I} - \mathcal{P}_{\mathcal{T}_1}]([\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2}](Y_2 - Z_2))\|_{\ell_2} \\ &\leq 2\sqrt{r}\|[\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2}](Y_2 - Z_2)\|_2 \\ &\leq 2\sqrt{r}\rho(\mathcal{T}_1, \mathcal{T}_2)\|Y_2 - Z_2\|_2 \\ &\leq \frac{4\sqrt{2}r}{\sigma}\frac{\delta_{4r}}{1-\delta_{4r}}\|X_1 - X_2\|_2\|Z_2\|_2.\end{aligned}$$

Here the first inequality follows by noting that $[\mathsf{I} - \mathcal{P}_{\mathcal{T}_1}]([\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2}](Y_2 - Z_2))$ has rank at most $4r$. Next

$$\|\mathcal{P}_{\mathcal{T}_1^\perp}(Z_2)\|_{\ell_2} = \|\mathcal{P}_{\mathcal{T}_1^\perp}(Z_1 - Z_2)\|_{\ell_2} \leq \|Z_1 - Z_2\|_{\ell_2} \leq 2\sqrt{r}\|Z_1 - Z_2\|_2.$$

By combining both bounds with the triangle inequality we obtain

$$\begin{aligned}\|\mathcal{P}_{\mathcal{T}_1^\perp}(Y_1 - Y_2)\|_{\ell_2} = \|\mathcal{P}_{\mathcal{T}_1^\perp}(Y_2)\|_{\ell_2} &\leq \|\mathcal{P}_{\mathcal{T}_1^\perp}(Z_2)\|_{\ell_2} + \|\mathcal{P}_{\mathcal{T}_1^\perp}(Y_2 - Z_2)\|_{\ell_2} \\ &\leq 2\sqrt{r}\|Z_1 - Z_2\|_2 + \frac{4\sqrt{2}r}{\sigma}\frac{\delta_{4r}}{1-\delta_{4r}}\|X_1 - X_2\|_2\|Z_2\|_2.\end{aligned}$$

$[\|\mathcal{P}_{\mathcal{T}_1}(Y_1 - Y_2)\|_{\ell_2}]$: Define the linear map $\mathsf{G} = \mathcal{L}'_{\mathcal{T}_1\cup\mathcal{T}_2}\mathcal{L}_{\mathcal{T}_1\cup\mathcal{T}_2}$. First $\|[\mathcal{P}_{\mathcal{T}_2} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2) - [\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2)\|_{\ell_2} \leq 2\sqrt{r}\|[\mathcal{P}_{\mathcal{T}_2} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2) - [\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2)\|_2 \leq 2\sqrt{r}\rho(\mathcal{T}_1, \mathcal{T}_2)\|\mathsf{G}(Y_2)\|_2$, where $\|\mathsf{G}(Y_2)\|_2 \leq \|\mathsf{G}(Y_2)\|_{\ell_2} \leq (1+\delta_{4r})\|Y_2\|_{\ell_2} \leq \frac{1+\delta_{4r}}{1-\delta_{4r}}\|Z_2\|_{\ell_2} \leq \sqrt{2r}\frac{1+\delta_{4r}}{1-\delta_{4r}}\|Z_2\|_2$. Second $\|[\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2) - [\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_1}](Y_2)\|_{\ell_2} = \|[\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ (\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2})](Y_2)\|_{\ell_2} \leq \|[\mathsf{G} \circ (\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2})](Y_2)\|_{\ell_2} \leq (1 + \delta_{4r})\|[\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2}](Y_2)\|_{\ell_2} \leq 2\sqrt{r}(1+\delta_{4r})\|[\mathcal{P}_{\mathcal{T}_1} - \mathcal{P}_{\mathcal{T}_2}](Y_2)\|_2 \leq 2\sqrt{r}(1+\delta_{4r})\rho(\mathcal{T}_1, \mathcal{T}_2)\|Y_2\|_2$, where $\|Y_2\|_2 \leq \|Y_2\|_{\ell_2} \leq \frac{\sqrt{2r}}{1-\delta_{4r}}\|Z\|_2$. Third by combining these bounds with an application of Lemma 2.3.9 and the triangle inequality we obtain

$$\begin{aligned}&\|\mathcal{P}_{\mathcal{T}_1}(Y_1 - Y_2)\|_{\ell_2} \\ \leq\ & \frac{1}{1-\delta_{4r}}\|[\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_1}](Y_1 - Y_2)\|_{\ell_2} \\ \leq\ & \frac{1}{1-\delta_{4r}}(\|[\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_1}](Y_1) - [\mathcal{P}_{\mathcal{T}_2} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2)\|_{\ell_2} \\ &+ \|[\mathcal{P}_{\mathcal{T}_2} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2) - [\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2)\|_{\ell_2} \\ &+ \|[\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_2}](Y_2) - [\mathcal{P}_{\mathcal{T}_1} \circ \mathsf{G} \circ \mathcal{P}_{\mathcal{T}_1}](Y_2)\|_{\ell_2}) \\ \leq\ & \frac{1}{1-\delta_{4r}}(2\sqrt{r}\|Z_1 - Z_2\|_2 + 4\sqrt{2}r\rho(\mathcal{T}_1, \mathcal{T}_2)\frac{1+\delta_{4r}}{1-\delta_{4r}}\|Z\|_2).\end{aligned}$$

□

Proof of Proposition A.6.1. We prove (i) and (ii) in sequence.

[(i)]: Let $\hat{X}_o$ be the optimal solution to the following

$$\hat{X}_o \;=\; \underset{X}{\operatorname{argmin}} \|\mathbf{y} - \mathcal{L}(X)\|_{\ell_2}^2 \qquad \text{s.t.} \qquad \operatorname{rank}(X) \le r, \qquad \|X - X^\star\|_2 \le 4\sqrt{r}\epsilon.$$

Since $4\sqrt{r}\epsilon < 1/2 \le \sigma_r(X^\star)$, $\hat{X}_o$ has rank exactly r, and hence is a smooth point with respect to the variety of matrices with rank at most r. Define the tangent space $\hat{\mathcal{T}} := \mathcal{T}(\hat{X}_o)$, and the matrix $\hat{X}_c$ as the solution to the following optimization instance

$$\hat{X}_c \;=\; \underset{X}{\operatorname{argmin}} \|\mathbf{y} - \mathcal{L}(X)\|_2^2 \qquad \text{s.t.} \qquad X \in \hat{\mathcal{T}}, \qquad \|X - X^\star\|_2 \le 4\sqrt{r}\epsilon.$$

Here $\hat{X}_c$ is the solution to the optimization instance where the constraint $X \in \hat{\mathcal{T}}$, which is convex, replaces the only non-convex constraint in the previous optimization instance. Hence $\hat{X}_c = \hat{X}_o$. Define $\hat{X}_{\hat{\mathcal{T}}}$ as the solution to the following optimization instance

$$\hat{X}_{\hat{\mathcal{T}}} \;=\; \underset{X}{\operatorname{argmin}} \|\mathbf{y} - \mathcal{L}(X)\|_{\ell_2}^2 \qquad \text{s.t.} \qquad X \in \hat{\mathcal{T}}.$$

The first order condition is given by $\mathcal{L}'\mathcal{L}(\hat{X}_{\hat{\mathcal{T}}} - X^\star) - \mathcal{L}'\mathbf{z} + Q_{\hat{\mathcal{T}}^\perp} = 0$, where $Q_{\hat{\mathcal{T}}^\perp} \in \hat{\mathcal{T}}^\perp$ is the Lagrange multiplier associated to the constraint $X \in \hat{\mathcal{T}}$. Project the above equation onto the subspace $\hat{\mathcal{T}}$ to obtain $[\mathcal{P}_{\hat{\mathcal{T}}} \circ \mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}}](\hat{X}_{\hat{\mathcal{T}}} - X^\star) = [\mathcal{P}_{\hat{\mathcal{T}}} \circ \mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}^\perp}](X^\star) + \mathcal{P}_{\hat{\mathcal{T}}}(\mathcal{L}'\mathbf{z})$, and hence

$$\hat{X}_{\hat{\mathcal{T}}} - X^\star = [(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \left([\mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}^\perp}](X^\star) + \mathcal{L}'\mathbf{z}\right) - \mathcal{P}_{\hat{\mathcal{T}}^\perp}(X^\star).$$

We proceed to bound $\|\hat{X}_{\hat{\mathcal{T}}} - X^\star\|_2$. First we have $\|\hat{X}_c - X^\star\|_2 \le 4\sqrt{r}\epsilon \le 1/20$, and hence $\sigma_r(\hat{X}_c) \ge 9/20$. Second by applying Proposition A.6.2, we have $\|\mathcal{P}_{\hat{\mathcal{T}}^\perp}(X^\star)\|_2 = \|\mathcal{P}_{\hat{\mathcal{T}}^\perp}(\hat{X}_c - X^\star)\|_2 \le (4\sqrt{r}\epsilon)^2/(5\sigma_r(\hat{X}_c)) \le (64/9)r\epsilon^2$, and $\|\mathcal{P}_{\hat{\mathcal{T}}^\perp}(X^\star)\|_{\ell_2} \le (320/27)r^{3/2}\epsilon^2$. Third by Lemma 2.3.9 and noting the inequality $\|\cdot\|_2 \le \|\cdot\|_{\ell_2}$ we have

$$\begin{aligned} \|[(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{p\times p}}(\mathcal{L}'\mathbf{z})\|_2 \;&\le\; \|[(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{p\times p}}\|_2 \|\mathcal{P}_{\hat{\mathcal{T}}}(\mathcal{L}'\mathbf{z})\|_{\ell_2} \\ &\le\; 2\sqrt{2r}\|\mathcal{L}'\mathbf{z}\|_2/(1-\delta_{4r}) \le (16/5)\sqrt{r}\epsilon. \end{aligned}$$

Fourth by Proposition 2.7 in [65] we have

$$\begin{aligned} &\|[[(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}^\perp}](X^\star)\|_2 \\ \le\; &\|[(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{p\times p}}\|_2 \|[\mathcal{P}_{\hat{\mathcal{T}}} \circ \mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}^\perp}](X^\star_{\hat{\mathcal{T}}^\perp})\|_{\ell_2} \\ \le\; &\delta_{4r}\|\mathcal{P}_{\hat{\mathcal{T}}^\perp}(X^\star)\|_{\ell_2}/(1-\delta_{4r}) \le (320/243)r^{3/2}\epsilon^2. \end{aligned}$$

Last, we combine the bounds to obtain $\|\hat{X}_{\hat{\mathcal{T}}} - X^\star\|_2 \leq 8r\epsilon^2 + (16/5)\sqrt{r}\epsilon + 2r^{3/2}\epsilon^2 < 4\sqrt{r}\epsilon$. This implies that the constraint $\|X - X^\star\|_2 \leq 4\sqrt{r}\epsilon$ for $\hat{X}_c$ and $\hat{X}_o$ are inactive, and hence $\hat{X} = \hat{X}_o = \hat{X}_c = \hat{X}_{\hat{\mathcal{T}}}$.

[(ii)]: We have

$$\begin{aligned} G &= [(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{p\times p}}(\mathcal{L}'\mathbf{z}) - [(\mathcal{L}'_{\mathcal{T}^\star}\mathcal{L}_{\mathcal{T}^\star})^{-1}]_{\mathbb{R}^{p\times p}}(\mathcal{L}'\mathbf{z}) \\ &+ [[(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}^\perp}](X^\star) - \mathcal{P}_{\hat{\mathcal{T}}^\perp}(X^\star). \end{aligned}$$

We deal with the contributions of each term separately.

First $\|[\mathcal{P}_{\mathcal{T}^\star} - \mathcal{P}_{\hat{\mathcal{T}}}](\mathcal{L}'\mathbf{z})\|_2 \leq \rho(\hat{\mathcal{T}}, \mathcal{T}^\star)\|\mathcal{L}'\mathbf{z}\|_2 \leq (2\epsilon/\sigma_r(X^\star))\|\hat{X} - X^\star\|_2 \leq 16\sqrt{r}\epsilon^2$, where the second inequality applies Proposition 2.1 of the s.m. of [34]. Second $\|\mathcal{P}_{\mathcal{T}^\star}(\mathcal{L}'\mathbf{z})\|_2 \leq 2\|\mathcal{L}'\mathbf{z}\|_2 = 2\epsilon$. Hence by applying Lemma A.6.3 with the choice of $Z_1 = \mathcal{P}_{\hat{\mathcal{T}}}(\mathcal{L}'\mathbf{z})$ and $Z_2 = \mathcal{P}_{\mathcal{T}^\star}(\mathcal{L}'\mathbf{z})$ we obtain $\|[(\mathcal{L}'_{\mathcal{T}^\star}\mathcal{L}_{\mathcal{T}^\star})^{-1}]_{\mathbb{R}^{p\times p}}(\mathcal{L}'\mathbf{z}) - [(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{p\times p}}(\mathcal{L}'\mathbf{z})\|_{\ell_2} \leq 70r\epsilon^2 + 256r^{3/2}\epsilon^2$. Third we have $\|[(\mathcal{L}'_{\hat{\mathcal{T}}}\mathcal{L}_{\hat{\mathcal{T}}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}'\mathcal{L} \circ \mathcal{P}_{\hat{\mathcal{T}}^\perp}](X^\star)\|_{\ell_2} \leq (320/243)r^{3/2}\epsilon^2$, and $\|\mathcal{P}_{\hat{\mathcal{T}}^\perp}(X^\star)\|_{\ell_2} \leq (320/27)r^{3/2}\epsilon^2$.

The bound follows by summing up these bounds. □

The proof of Proposition 2.3.2 requires two additional preliminary results; in particular, the first establishes the restricted isometry condition for linear maps that are near linear maps that already satisfy the restricted isometry condition.

Proposition A.6.4 *Suppose $\mathcal{L}^\star$ is a linear map that satisfies the restricted isometry condition $\delta_r(\mathcal{L}^\star) \leq 1/20$. Let E be a linear operator such that $\|\mathsf{E}\|_2 \leq 1/(50\|\mathcal{L}^\star\|_2)$. Then $\mathcal{L} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E})$ satisfies the restricted isometry condition $\delta_r(\mathcal{L}) \leq 1/10$.*

Proof of Proposition A.6.4. Let X be a matrix with rank at most r. Then

$$\|\mathcal{L}(X)\|_{\ell_2} \leq \|\mathcal{L}^\star(X)\|_{\ell_2} + \|\mathcal{L}^\star(\mathsf{E}(X))\|_{\ell_2} \leq (\sqrt{1+\delta_r(\mathcal{L}^\star)} + 1/50)\|X\|_{\ell_2} \leq \sqrt{1+1/10}\|X\|_{\ell_2}.$$

A similar argument also proves the lower bound $\|\mathcal{L}(X)\|_{\ell_2} \geq \sqrt{1-1/10}\|X\|_{\ell_2}$. □

Lemma A.6.5 *Suppose $\mathcal{L}$ satisfies the restricted isometry condition $\delta_1(\mathcal{L}) < 1$. Then $\|\mathcal{L}'\mathcal{L}\|_{\ell_2,2} \leq \sqrt{2(1+\delta_1(\mathcal{L}))}\|\mathcal{L}\|_2$.*

Proof. Let $Z \in \operatorname{argmax}_{X:\|X\|_{\ell_2}\leq 1}\|\mathcal{L}'\mathcal{L}(X)\|_2$, and let $\mathcal{T}$ be the tangent space of the rank-one matrix corresponding to the largest singular value of Z. Then $\sup_{X:\|X\|_{\ell_2}\leq 1}\|\mathcal{L}'\mathcal{L}(X)\|_2 \leq \sup_{X:\|X\|_{\ell_2}\leq 1}\|[\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}](X)\|_2 \leq \sqrt{2}\sup_{X:\|X\|_{\ell_2}\leq 1}\|[\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}](X)\|_{\ell_2} \leq \sqrt{2}\|\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}\|_2$. By Lemma 2.3.9 we have $\sqrt{2}\|\mathcal{P}_{\mathcal{T}} \circ \mathcal{L}'\mathcal{L}\|_2 \leq \sqrt{2(1+\delta_1(\mathcal{L}))}\|\mathcal{L}\|_2$. □

Proof of Proposition 2.3.2. To simplify notation we denote $\mathcal{T} := \mathcal{T}(X^\star)$. Without loss of generality we may assume that $\|X^\star\|_2 = 1$. By the triangle inequality we have

$$\begin{aligned}
&\|(X^\star - \mathcal{M}(\hat{X})) - [(\mathcal{L}^{\star\prime}_{\mathcal{T}}\mathcal{L}^\star_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star \circ \mathsf{E}(X^\star)\|_{\ell_2} \\
\leq\ &\|(X^\star - \mathcal{M}(\hat{X})) - [((\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ(\mathsf{I}+\mathsf{E})|_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}} \circ (\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ\mathsf{E}(X^\star)\|_{\ell_2} \\
+\ &\|([((\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ(\mathsf{I}+\mathsf{E})|_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}} - [(\mathcal{L}^{\star\prime}_{\mathcal{T}}\mathcal{L}^\star_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}}) \circ (\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ\mathsf{E}(X^\star)\|_{\ell_2} \\
+\ &\|[(\mathcal{L}^{\star\prime}_{\mathcal{T}}\mathcal{L}^\star_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}} \circ (\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ\mathsf{E}(X^\star) - [(\mathcal{L}^{\star\prime}_{\mathcal{T}}\mathcal{L}^\star_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star\circ\mathsf{E}(X^\star)\|_{\ell_2}
\end{aligned}$$

We bound each term separately.

[First term]: Let $\tilde{\mathbf{z}} := [\mathcal{L}^\star \circ \mathsf{E}](X^\star)$. First by Proposition A.6.4 the linear map $\mathcal{L}^\star \circ (\mathsf{I}+\mathsf{E})$ satisfies the restricted isometry condition $\delta_{4r}(\mathcal{L}^\star\circ(\mathsf{I}+\mathsf{E})) \leq 1/10$. Second we have $\|\mathsf{I}+\mathsf{E}\|_{2,2} \leq 1 + \sqrt{p}\|\mathsf{E}\|_{\ell_2} \leq 51/50$. Third from Lemma A.6.5 we have $\|\mathcal{L}^{\star\prime}\mathcal{L}^\star\|_{\ell_2,2} \leq \sqrt{2(1+\delta_{4r}(\mathcal{L}^\star))}\|\mathcal{L}^\star\|_2$. Fourth $\|\mathsf{E}(X^\star)\|_{\ell_2} \leq \sqrt{r}\|\mathsf{E}\|_{\ell_2}$. Hence

$$\|(\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\tilde{\mathbf{z}}\|_2 \leq \|\mathsf{I}+\mathsf{E}\|_{2,2}\|\mathcal{L}^{\star\prime}\mathcal{L}^\star\|_{\ell_2,2}\|\mathsf{E}\|_{\ell_2}\|X^\star\|_{\ell_2} \leq (3/2)\sqrt{r}\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2}.$$

By the initial conditions we have that the above quantity is at most $1/(80r^{3/2})$. Consequently, by applying Proposition A.6.1 to the optimization instance (2.9) with the choice of linear map $\mathcal{L}^\star\circ(\mathsf{I}+\mathsf{E})$ and error term $\tilde{\mathbf{z}}$ we have

$$\begin{aligned}
&\|(X^\star - \mathcal{M}(\hat{X})) - [((\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ(\mathsf{I}+\mathsf{E})|_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}} \circ (\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\tilde{\mathbf{z}}\|_{\ell_2} \\
\leq\ &765r^{5/2}\|\mathcal{L}^\star\|_2^2\|\mathsf{E}\|_{\ell_2}^2.
\end{aligned}$$

[Second term]: First by Lemma 2.3.9 we have $\|[(\mathcal{L}^{\star\prime}_{\mathcal{T}}\mathcal{L}^\star_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}}\|_2 \leq 20/19$. Second by the triangle inequality we have $\|\mathcal{P}_{\mathcal{T}}\circ(\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ(\mathsf{I}+\mathsf{E})\circ\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ\mathcal{P}_{\mathcal{T}}\|_2 \leq 3\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2}$. Third by utilizing the identity $(\mathsf{A}+\mathsf{B})^{-1} = \mathsf{A}^{-1} - \mathsf{A}^{-1}\circ\mathsf{B}\circ\mathsf{A}^{-1} + \mathsf{A}^{-1}\circ\mathsf{B}\circ\mathsf{A}^{-1}\circ\mathsf{B}\circ\mathsf{A}^{-1} - \ldots$ with the choice of $\mathsf{A} = \mathcal{L}^{\star\prime}_{\mathcal{T}}\mathcal{L}^\star_{\mathcal{T}}$ and $\mathsf{B} = \mathcal{P}_{\mathcal{T}}\circ(\mathsf{I}+\mathsf{E})'\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ(\mathsf{I}+\mathsf{E})\circ\mathcal{P}_{\mathcal{T}} - \mathsf{A}$ we obtain

$$\|[((\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ(\mathsf{I}+\mathsf{E})|_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}} - [(\mathcal{L}^{\star\prime}_{\mathcal{T}}\mathcal{L}^\star_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}}\|_2 \leq 4\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2}.$$

Fourth $\|\mathcal{P}_{\mathcal{T}}\circ(\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ\mathsf{E}(X^\star)\|_{\ell_2} \leq (11/10)\sqrt{r}\|\mathcal{L}^\star\|_2\|\mathsf{E}\|_{\ell_2}$. Hence

$$\begin{aligned}
&\|([((\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ(\mathsf{I}+\mathsf{E})|_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}} - [(\mathcal{L}^{\star\prime}_{\mathcal{T}}\mathcal{L}^\star_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}}) \circ (\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ\mathsf{E}(X^\star)\|_{\ell_2} \\
\leq\ &5\sqrt{r}\|\mathcal{L}^\star\|_2^2\|\mathsf{E}\|_{\ell_2}^2.
\end{aligned}$$

[Third Term]: We have

$$\begin{aligned}
&\|[(\mathcal{L}^{\star\prime}_{\mathcal{T}}\mathcal{L}^\star_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}} \circ (\mathsf{I}+\mathsf{E}')\circ\mathcal{L}^{\star\prime}\mathcal{L}^\star\circ\mathsf{E}(X^\star) - [(\mathcal{L}^{\star\prime}_{\mathcal{T}}\mathcal{L}^\star_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}} \circ \mathcal{L}^{\star\prime}\mathcal{L}^\star\circ\mathsf{E}(X^\star)\|_{\ell_2} \\
\leq\ &\|[(\mathcal{L}^{\star\prime}_{\mathcal{T}}\mathcal{L}^\star_{\mathcal{T}})^{-1}]_{\mathbb{R}^{p\times p}}\|_2\|\mathsf{E}'\|_2\|\mathcal{L}^\star\|_2^2\|\mathsf{E}(X^\star)\|_{\ell_2} \leq 2\sqrt{r}\|\mathcal{L}^\star\|_2^2\|\mathsf{E}\|_{\ell_2}^2.
\end{aligned}$$

[Conclude]: The result follows by summing each bound and applying Lemma A.6.5. □

A.7 Proof of Proposition 2.3.3

Proof of Proposition 2.3.3. To simplify notation we let $\Lambda := \Lambda(\{A^{(j)}\}_{j=1}^n)$, $\Delta := \Delta(\{A^{(j)}\}_{j=1}^n)$, and $\mathtt{D}$ be the linear map defined as $\mathtt{D} : \mathbf{z} \mapsto \sum_{j=1}^n (\mathsf{Q}(B^{(j)}) - A^{(j)})\mathbf{z}_j$. In addition we define $\tau := (1/\sqrt{n\Lambda})\|\mathtt{D}\|_2$. Note that by the Cauchy-Schwarz inequality we have $\tau \leq \omega/\sqrt{\Lambda} \leq 1/20$.

We begin by noting that since $\|(1/n\Lambda)\mathtt{X}^\star \circ \mathtt{X}^{\star\prime} - \mathsf{I}\|_2 \leq \Delta/\Lambda \leq 1/6$, we have $\|((1/n\Lambda)\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1}\|_2$, and $\|(1/n\Lambda)\mathtt{X}^\star \circ \mathtt{X}^{\star\prime}\|_2 \leq 6/5$.

Next we compute the following bounds. First $\|\mathtt{D} \circ \mathtt{D}' \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1}\|_2 \leq \|\mathtt{D}\|_2^2 \|(\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1}\|_2 \leq (6/5)\tau^2$. Second $\|\mathtt{D} \circ \mathtt{X}^{\star\prime} \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1}\|_2 \leq \|\mathtt{D}\|_2 \|\mathtt{X}^{\star\prime}\|_2 \|(\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1}\|_2 \leq \tau(6/5)^{3/2}$. Third $\|\mathtt{X}^\star \circ \mathtt{D}' \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1}\|_2 \leq \tau(6/5)^{3/2}$. By applying these bounds to the following expansion we obtain

$$\begin{aligned}
&\big((\mathtt{X}^\star + \mathtt{D}) \circ (\mathtt{X}^\star + \mathtt{D})'\big)^{-1} \\
&= \big(\big(\mathsf{I} + \mathtt{D} \circ \mathtt{X}^{\star\prime} \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1} + \mathtt{X}^\star \circ \mathtt{D}' \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1} + \mathsf{E}_1\big) \circ \mathtt{X}^\star \circ \mathtt{X}^{\star\prime}\big)^{-1} \\
&= (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1}\big(\mathsf{I} - \mathtt{D} \circ \mathtt{X}^{\star\prime} \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1} - \mathtt{X}^\star \circ \mathtt{D}' \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1} + \mathsf{E}_2\big),
\end{aligned}$$

where $\|\mathsf{E}_1\|_2 \leq (6/5)\tau^2$, and $\|\mathsf{E}_2\|_2 = \| - \mathsf{E}_1 + (\mathtt{D} \circ \mathtt{X}^{\star\prime} \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1} + \mathtt{X}^\star \circ \mathtt{D}' \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1} + \mathsf{E}_1)^2 - (\ldots)^3\|_2 \leq (\|\mathsf{E}_1\|_2 + \|\mathtt{D} \circ \mathtt{X}^{\star\prime} \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1} + \mathtt{X}^\star \circ \mathtt{D}' \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1} + \mathsf{E}_1\|_2^2 + \ldots) \leq (6/5)\tau^2 + (\tau(6/5)(\tau + 2\sqrt{6/5}))^2 + \ldots \leq 1.2\tau^2 + (3\tau)^2 + (3\tau)^3 + \ldots \leq 12\tau^2$.

We apply the above expansion to derive the following approximation of $\mathtt{X}^\star \circ (\mathtt{X}^\star + \mathtt{D})^+$

$$\begin{aligned}
&\mathtt{X}^\star \circ (\mathtt{X}^\star + \mathtt{D})^+ \\
&= \mathtt{X}^\star \circ (\mathtt{X}^\star + \mathtt{D})' \circ \big((\mathtt{X}^\star + \mathtt{D}) \circ (\mathtt{X}^\star + \mathtt{D})'\big)^{-1} \\
&= (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime} + \mathtt{X}^\star \circ \mathtt{D}') \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1}\big(\mathsf{I} - \mathtt{D} \circ \mathtt{X}^{\star\prime} \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1} - \mathtt{X}^\star \circ \mathtt{D}' \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1} + \mathsf{E}_2\big) \\
&= (\mathsf{I} - \mathtt{D} \circ \mathtt{X}^{\star+} + \mathsf{E}_3),
\end{aligned}$$

where E_3 satisfies $\|\mathsf{E}_3\|_2 \leq 2(\tau(6/5)^{3/2})(2(\tau(6/5)^{3/2}) + \|\mathsf{E}_2\|_2) + \|\mathsf{E}_2\|_2 \leq 20\tau^2$.

Next we write $((1/(n\Lambda))\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1} = \mathsf{I} + \mathsf{E}_4$, where $\|\mathsf{E}_4\|_2 \leq (6/5)\Delta/\Lambda$. Then

$$\mathtt{X}^\star \circ (\mathtt{X}^\star + \mathtt{D})^+ = \mathsf{I} - \mathtt{D} \circ \mathtt{X}^{\star\prime} \circ (\mathtt{X}^\star \circ \mathtt{X}^{\star\prime})^{-1} + \mathsf{E}_3 = \mathsf{I} - (1/n\Lambda)\mathtt{D} \circ \mathtt{X}^{\star\prime} + \mathsf{F},$$

where $\|\mathsf{F}\|_2 \leq \|\mathsf{E}_3\|_2 + \|\mathtt{D} \circ \mathtt{X}^{\star\prime} \circ \mathsf{E}_4\|_2/(n\Lambda) \leq \|\mathsf{E}_3\|_2 + \tau(6/5)^{1/2}\|\mathsf{E}_4\|_2 \leq 20\tau^2 + 2\tau\Delta/\Lambda$. The result follows by noting that $\|\mathsf{F}\|_{\ell_2} \leq q\|\mathsf{F}\|_2$, $\tau \leq \omega/\sqrt{\Lambda}$, and that $\mathtt{X}^\star \circ \hat{\mathtt{X}}^+ = \mathtt{X}^\star \circ (\mathtt{X}^\star + \mathtt{D})^+ \circ \mathsf{Q}$. □

A.8 Proof of Proposition 2.3.4

Proposition A.8.1 *Given an operator* $\mathsf{E} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$*, there exists matrices* E_L*,* E_R *such that* $\mathcal{P}_{\mathcal{W}}(\mathsf{E}) = I \otimes E_L + E_R \otimes I$*, and* $\|E_L\|_{\ell_2}, \|E_R\|_{\ell_2} \le \|\mathsf{E}\|_{\ell_2}/\sqrt{p}$.

Proof of Proposition A.8.1. Define the subspaces $\mathcal{W}_R := \{S \otimes I : S \in \mathbb{R}^{p\times p}\}$ and $\mathcal{W}_L := \{I \otimes S : S \in \mathbb{R}^{p\times p}\}$. Note that $\mathcal{W}_R \cap \mathcal{W}_L = \mathrm{Span}(\mathsf{I})$, and hence $\mathcal{P}_{\mathcal{W}} = \mathcal{P}_{\mathcal{W}_R \cap \mathrm{Span}(\mathsf{I})^\perp} + \mathcal{P}_{\mathcal{W}_L \cap \mathrm{Span}(\mathsf{I})^\perp} + \mathcal{P}_{\mathrm{Span}(\mathsf{I})}$.

Define E_L and E_R to be matrices such that $E_R \otimes I = \mathcal{P}_{\mathcal{W}_R \cap \mathrm{Span}(\mathsf{I})^\perp}(\mathsf{E}) + (1/2)\mathcal{P}_{\mathrm{Span}(\mathsf{I})}(\mathsf{E})$, and $I \otimes E_L = \mathcal{P}_{\mathcal{W}_L \cap \mathrm{Span}(\mathsf{I})^\perp}(\mathsf{E}) + (1/2)\mathcal{P}_{\mathrm{Span}(\mathsf{I})}(\mathsf{E})$. For $i \in \{L, R\}$ we have the following. Since $\mathcal{P}_{\mathcal{W}_i \cap \mathrm{Span}(\mathsf{I})^\perp}$ and $(1/2)\mathcal{P}_{\mathrm{Span}(\mathsf{I})}$ are projectors onto orthogonal subspaces with spectral norm 1 and $1/2$ respectively, we have $\|E_i \otimes I\|_{\ell_2} \le \|\mathsf{E}\|_{\ell_2}$. Moreover, since $\|E_i \otimes I\|_{\ell_2} = \|E_i\|_{\ell_2}\|I\|_{\ell_2}$, we have $\|E_i\|_{\ell_2} \le \|\mathsf{E}\|_{\ell_2}/\sqrt{p}$. □

Proof of Proposition 2.3.4. By applying Proposition A.8.1 to the operator D we have $\mathcal{P}_{\mathcal{W}}(\mathsf{D}) = I \otimes E_L + E_R \otimes I$ for a pair of matrices $E_L, E_R \in \mathbb{R}^{p\times p}$ satisfying $\|E_L\|_{\ell_2}, \|E_R\|_{\ell_2} \le \|\mathsf{D}\|_{\ell_2}/\sqrt{p}$. Moreover since $\|E_L\|_2, \|E_R\|_2 < 1$, it follows that the matrices $I + E_R$ and $I + E_L$ are invertible. Consider the following identity

$$\mathsf{I} + \mathsf{D} = \Big(\mathsf{I} + (\mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D}) - E_R \otimes E_L) \circ (I + E_R)^{-1} \otimes (I + E_L)^{-1}\Big) \circ (I + E_R) \otimes (I + E_L).$$

We define $\mathsf{H} = (\mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D}) - E_R \otimes E_L) \circ (I + E_R)^{-1} \otimes (I + E_L)^{-1} - \mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D})$, and we define $\mathsf{W} = (I + E_R) \otimes (I + E_L)$. By the triangle inequality we have $\|\mathsf{W} - \mathsf{I}\|_2 \le 3\|\mathsf{D}\|_{\ell_2}/\sqrt{p}$.

Next we note that $\|(I + E_i)^{-1}\|_2 \le 10/9$, $i \in \{L, R\}$, and that $\|(I + E_R)^{-1} \otimes (I + E_L)^{-1}\|_2 \le 100/81$. We also have $\|E_R \otimes E_L\|_{\ell_2} = \|E_R\|_{\ell_2}\|E_L\|_{\ell_2} \le \|\mathsf{D}\|_{\ell_2}^2/q$. By noting that $\|(I + E_i)^{-1} - I\|_2 \le (10/9)\|E_i\|_2$, $i \in \{L, R\}$, we have $\|(I + E_R)^{-1} \otimes (I + E_L)^{-1} - I \otimes I\|_2 \le 3\|\mathsf{D}\|_{\ell_2}/\sqrt{p}$. By combining these bounds we obtain $\|\mathsf{H}\|_{\ell_2} \le \|\mathcal{P}_{\mathcal{W}^\perp}(\mathsf{D})\|_{\ell_2}\|(I+E_R)^{-1}\otimes(I+E_L)^{-1} - I\otimes I\|_2 + \|E_R\otimes E_L\|_{\ell_2}\|(I+E_R)^{-1}\otimes(I+E_L)^{-1}\|_2 \le 5\|\mathsf{D}\|_{\ell_2}^2/\sqrt{p}$. □

A.9 Proof of Proposition 2.3.6

Proof of Proposition 2.3.6. To simplify notation in the proof we denote $\alpha_8 := \alpha_8(p, \mathcal{L}^\star) = 96\sqrt{p}\|\mathcal{L}^\star\|_2$. We show that

$$\|\mathcal{L}^{(t)} - \mathcal{L}^{(t+1)}\|_2 \le \alpha_9 \xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)}), \tag{A.5}$$

for some function $\alpha_9 := \alpha_9(p, r, \mathcal{L}^\star)$ that we specify later. In the proof of Theorem 2.3.5 we showed that $\xi_{\mathcal{L}^\star}(\mathcal{L}^{(t)}) \le \gamma^t \xi_{\mathcal{L}^\star}(\mathcal{L}^{(0)})$ for some $\gamma < 1$. Hence establishing (A.5) immediately implies that the sequence $\{\mathcal{L}^{(t)}\}_{t=1}^{\infty}$ is Cauchy.

Our proof builds on the proof of Theorem 2.3.5. Let

$$\mathcal{L}^{(t)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t)}) \circ \mathsf{M}$$

where $\mathsf{E}^{(t)}$ is a linear map that satisfies $\|\mathsf{E}^{(t)}\|_{\ell_2} < 1/\alpha_0$. In the proof of Theorem 2.3.5 we show that

$$\mathcal{L}^{(t+1)} = \mathcal{L}^\star \circ (\mathsf{I} + \mathsf{E}^{(t+1)}) \circ \mathsf{W} \circ \mathsf{M} \circ \mathsf{N},$$

where $\|\mathsf{E}^{(t+1)}\|_{\ell_2} \leq \|\mathsf{E}^{(t)}\|_{\ell_2}$, W is a rank-preserver, and N is a positive definite rank-preserver. Moreover, as a consequence of applying Proposition 2.3.4 to establish (2.23) in the proof, we obtain the bound $\|\mathsf{W} - \mathsf{I}\|_2 \leq 3\alpha_7\|\mathsf{E}^{(t)}\|_{\ell_2}$. We use these bounds and relations to prove (A.5).

By the triangle inequality we have

$$\begin{aligned} \|\mathcal{L}^{(t)} - \mathcal{L}^{(t+1)}\|_2 &\leq \|\mathcal{L}^\star \circ \mathsf{E}^{(t)} \circ \mathsf{M}\|_2 + \|\mathcal{L}^\star \circ \mathsf{E}^{(t+1)} \circ \mathsf{W} \circ \mathsf{M} \circ \mathsf{N}\|_2 \\ &+ \|\mathcal{L}^\star \circ \mathsf{M} \circ (\mathsf{N} - \mathsf{I})\|_2 + \|\mathcal{L}^\star \circ (\mathsf{W} - \mathsf{I}) \circ \mathsf{M} \circ \mathsf{N}\|_2. \quad \text{(A.6)} \end{aligned}$$

By Proposition 2.3.1 applied to the pairs of linear maps $\mathcal{L}^{(t)}, \mathcal{L}^\star$ and $\mathcal{L}^{(t+1)}, \mathcal{L}^\star$ we have $\|\mathsf{M} - \mathsf{Q}_1\|_2, \|\mathsf{W} \circ \mathsf{M} \circ \mathsf{N} - \mathsf{Q}_2\|_2 \leq \alpha_8\|\mathsf{E}^{(t)}\|_{\ell_2}$, for some pair of orthogonal rank-preservers $\mathsf{Q}_1, \mathsf{Q}_2$. Since $\alpha_8/\alpha_0 \leq 1$ we have $\|\mathsf{M}\|_2 \leq 2$ and $\|\mathsf{W} \circ \mathsf{M} \circ \mathsf{N}\|_2 \leq 2$. Consequently $\|\mathcal{L}^\star \circ \mathsf{E}^{(t)} \circ \mathsf{M}\|_2, \|\mathcal{L}^\star \circ \mathsf{E}^{(t+1)} \circ \mathsf{W} \circ \mathsf{M} \circ \mathsf{N}\|_2 \leq 2\|\mathcal{L}^\star\|_2\|\mathsf{E}^{(t)}\|_{\ell_2}$.

Next we bound $\|\mathsf{N} - \mathsf{I}\|_2$. By utilizing $\|\mathsf{W} \circ \mathsf{M} \circ \mathsf{N} - \mathsf{Q}_2\|_2 \leq \alpha_8/\alpha_0$, $\|\mathsf{M} - \mathsf{Q}_1\|_2 \leq \alpha_8\|\mathsf{E}^{(t)}\|_{\ell_2}$, and $\|\mathsf{W} - \mathsf{I}\|_2 \leq 3\alpha_7\|\mathsf{E}^{(t)}\|_{\ell_2}$, one can show that $\|\mathsf{N} - \mathsf{Q}_3\|_2 \leq (6\alpha_7 + 2\alpha_8 + 2)\|\mathsf{E}^{(t)}\|_{\ell_2}$, where $\mathsf{Q}_3 = \mathsf{Q}_1' \circ \mathsf{Q}_2$ is an orthogonal rank-preserver. Since N is self-adjoint, we have $\|\mathsf{N}^2 - \mathsf{I}\|_2 \leq 3(6\alpha_7 + 2\alpha_8 + 2)\|\mathsf{E}^{(t)}\|_{\ell_2}$, and hence $\|\mathsf{N} - \mathsf{I}\|_2 \leq 3(6\alpha_7 + 2\alpha_8 + 2)\|\mathsf{E}^{(t)}\|_{\ell_2}$. This also implies the bound $\|\mathsf{N}\|_2 \leq 3$.

We apply these bounds to obtain $\|\mathcal{L}^\star \circ \mathsf{M} \circ (\mathsf{N} - \mathsf{I})\|_2 \leq 6(6\alpha_7 + 2\alpha_8 + 2)\|\mathcal{L}^\star\|_2\|\mathsf{E}^{(t)}\|_{\ell_2}$, and $\|\mathcal{L}^\star \circ (\mathsf{W} - \mathsf{I}) \circ \mathsf{M} \circ \mathsf{N}\|_2 \leq 9\alpha_7\|\mathcal{L}^\star\|_2\|\mathsf{E}^{(t)}\|_2$.

We define $\alpha_9 := (4 + 6(6\alpha_7 + 2\alpha_8 + 2) + 9\alpha_7)\|\mathcal{L}^\star\|_2$ (these are exactly the sum of the coefficients of $\|\mathsf{E}^{(t)}\|_{\ell_2}$ in the above bounds). The result follows by adding these bounds, and subsequently taking the infimum over $\mathsf{E}^{(t)}$ in (A.6). □

Appendix B

PROOFS OF CHAPTER 3

We begin by noting the following bound, which we use subsequently in the proof of Proposition 3.2.1 and Lemma 3.3.2. For any pair of linear maps $A_1, A_2 \in L(\mathbb{R}^q, \mathbb{R}^d)$, any unit-norm vector $\mathbf{u}$, and any scalar y, we have $||h_C(A_1'\mathbf{u}) - y| - |h_C(A_2'\mathbf{u}) - y|| \leq |h_C(A_1'\mathbf{u}) - h_C(A_2'\mathbf{u})| \leq \|A_1 - A_2\|_{C,2}$.

Proof of Proposition 3.2.1. Let $r = 1 + \|A\|_{C,2}$. Let $\epsilon > 0$ be arbitrary, and let $\delta = \min\{\epsilon/(3\mathbb{E}\{r + |y|\}), r\}$. Then for any A_0 satisfying $\|A - A_0\|_{C,2} < \delta$, we have $|\Phi(A, Q) - \Phi(A_0, Q)| = |\mathbb{E}_Q\{(h_C(A'\mathbf{u}) - y)^2 - (h_C(A_0'\mathbf{u}) - y)^2\}| \leq \mathbb{E}_Q\{|h_C(A'\mathbf{u}) - h_C(A_0'\mathbf{u})|(2|h_C(A'\mathbf{u})| + 2|y| + \|A_0 - A\|_{C,2})\} \leq \delta\mathbb{E}_Q\{2r + \delta + 2|y|\} \leq \epsilon$. □

Proof of Proposition 3.2.2. Pick any $c \in (0, 1)$. Define $\mathcal{G}_{\mathbf{v},r,c} := \{(\mathbf{u}, y) : \langle \mathbf{v}, \mathbf{u}\rangle \geq c, |y| \leq rc/2\}$, for all $\mathbf{v} \in \mathcal{S}^{d-1}$, $r \geq 0$. Using a sequence of steps that is identical to the proof in Theorem 3.3.1, one can pick r sufficiently large so that $\mathbb{P}\{\mathcal{G}_{\mathbf{v},r,c}\}r^2c^2/4 > \Phi_C(0, P_{\mathcal{K}^\star})$ uniformly over all $\mathbf{v} \in \mathcal{S}^{d-1}$ (we omit the justification).

Suppose $A \notin r\mathcal{B}_{\|\cdot\|_{C,2}}(0)$. Then there exists $\mathbf{x} \in C$ such that $\|A\mathbf{x}\|_2 > r$. Define $\mathbf{v} = A\mathbf{x}/\|A\mathbf{x}\|_2$. Then

$$\Phi_C(A, P_{\mathcal{K}^\star}) \geq \mathbb{E}\{\mathbf{1}(\mathcal{G}_{\mathbf{v},r,c})(h_C(A'\mathbf{u}) - y)^2\} \geq \mathbb{P}\{\mathcal{G}_{\mathbf{v},r,c}\}r^2c^2/4 > \Phi_C(0, P_{\mathcal{K}^\star}),$$

and hence $A \notin \mathcal{M}_{\mathcal{K}^\star,C}$. This implies that $\mathcal{M}_{\mathcal{K}^\star,C} \subset r\mathcal{B}_{\|\cdot\|_{C,2}}(0)$, and thus $\mathcal{M}_{\mathcal{K}^\star,C}$ is bounded.

By Proposition 3.2.1, the function $A \mapsto \Phi_C(A, P_{\mathcal{K}^\star})$ is continuous. Since $r\mathcal{B}_{\|\cdot\|_{C,2}}(0)$ is compact, it follows that the minimizers of $\Phi_C(\cdot, P_{\mathcal{K}^\star})$ are attained, i.e., $\mathcal{M}_{\mathcal{K}^\star,C}$ is non-empty. By the continuity of $\Phi_C(\cdot, P_{\mathcal{K}^\star})$, it follows that $\mathcal{M}_{\mathcal{K}^\star,C}$ is closed, and hence also compact.

By Fubini's theorem, we have $\mathbb{E}\{\varepsilon(h_C(\mathcal{K}^\star) - h_C(A'\mathbf{u}))\} = \mathbb{E}_{\mathbf{u}}[\mathbb{E}_\varepsilon\{\varepsilon(h_C(\mathcal{K}^\star) - h_C(A'\mathbf{u}))\}] = 0$. Hence $\Phi_C(A, P_{\mathcal{K}^\star}) = \mathbb{E}\{(h_C(\mathcal{K}^\star) + \varepsilon - h_C(A'\mathbf{u}))^2\} = \mathbb{E}\{(h_C(\mathcal{K}^\star) - h_C(A'\mathbf{u}))^2\} + \mathbb{E}\{\varepsilon^2\}$, from which the last assertion follows. □

Proof of Corollary 3.2.3. Clearly $A^\star \in \mathcal{M}_{\mathcal{K}^\star,C}$. Since, for every $A \in L(\mathbb{R}^q, \mathbb{R}^d)$, $h_C(A'\mathbf{u})$ is a continuous function of $\mathbf{u}$ over a compact domain, it follows that

$\hat{A} \in \mathcal{M}_{\mathcal{K}^\star,C}$ if and only if $h_C(A^{\star\prime}\mathbf{u}) = h_C(\hat{A}'\mathbf{u})$ everywhere. By applying Proposition 3.2.2 and using the fact that a pair of compact convex sets that have the same support function must be equal, it follows that $\mathcal{K}^\star = \hat{A}(C)$ for all $\hat{A} \in \mathcal{M}_{\mathcal{K}^\star,C}$. □

Proof of Proposition 3.2.4. It is clear that $\mathcal{K}$ is expressible as the projection of $\mathcal{O}^q$. It remains to check that if A_1 and A_2 are linear maps such that $A_1(\mathcal{O}^q) = A_2(\mathcal{O}^q) = \mathcal{K}^\star$, then $A_1 = A_2 g$ for some $g \in \mathrm{Aut}(\mathcal{O}^q)$.

Let $\mathcal{F}_{i,j}$ be the pre-image of $\mathcal{K}_j$ under the map A_i. Since $\mathcal{K}_j$ is an exposed face, the pre-image $\mathcal{F}_{i,j}$ must be disjoint faces of $\mathcal{O}^q$, for each i. As such, the set $\mathcal{F}_{i,j}$ is isormorphic to the free spectrahedron $\mathcal{O}^{\tilde{q}_i}$, for some $\tilde{q}_i \leq q$. It is easy to check that $\tilde{q}_i = q_i$. Then the faces $\mathcal{F}_{i,j}$ are block diagonal matrices of $\mathbb{S}^q$ under a suitable choice of basis. Last, by using the fact that $\mathcal{M}_{\mathcal{K}_i,\mathcal{O}^{q_i}}$ is a single orbit, one can easily read off an element $g \in \mathrm{Aut}(\mathcal{O}^q)$ such that $A_1 = A_2 g$. □

Proof of Proposition 3.2.5. First, we have

$$\begin{aligned}
&|\lambda_C(\cdot, A, D)| \\
\leq\ & \|2(h_C(A'\mathbf{u}) - y)\mathbf{u} \otimes \boldsymbol{e}_C(A'\mathbf{u})\|_{C,2} \\
&\quad + |h_C((A+D)'\mathbf{u}) - h_C(A'\mathbf{u})||h_C((A+D)'\mathbf{u}) + h_C(A'\mathbf{u}) - 2y|/\|D\|_{C,2} \\
\leq\ & c_A(1 + |y|),
\end{aligned}$$

where c_A is a constant depending only on A. By noting that $\mathbb{E}_Q\{y^2\} < \infty$, we have $\lambda_C((\mathbf{u}, y), A, D) \in \mathcal{L}^2(Q)$. Second, since the function $h_C(\cdot)$ is differentiable at $A'\mathbf{u}$ for Q-a.e., we have

$$(h_C((A+D)'\mathbf{u}) - y)^2 = (h_C(A'\mathbf{u}) - y)^2 + \langle \nabla_A((h_C(A'\mathbf{u}) - y)^2), D\rangle + \lambda_C(\cdot, A, D)\|D\|_{C,2},$$

where $\lambda_C(\cdot, A, D) \to 0$ as $\|D\|_{C,2} \to 0$, for P-a.e. $\mathbf{u}$. Since $\lambda_C(\cdot, A, D) \in \mathcal{L}^2(Q)$, we also have $\lambda_C(\cdot, A, D) \in \mathcal{L}^1(Q)$. The second assertion follows from an application of the Dominated Convergence Theorem. □

Proof of Lemma 3.3.2. By a result in Section 1.1 of [58], it suffices to construct a sequence of finite function classes $\{\mathfrak{G}_\epsilon\}_{\epsilon>0}$ with the property that, for every $g \in \mathfrak{G}$, there is a pair $\overline{g}, \underline{g} \in \mathfrak{G}_\epsilon$ satisfying (i) $\underline{g} \leq \overline{g}$, and (ii) $\mathbb{E}_Q\{\overline{g} - \underline{g}\} < \epsilon$.

Our construction of $\mathfrak{G}_\epsilon$ is as follows. Without loss of generality assume that $\mathcal{U} \in r\mathcal{B}_{\|\cdot\|_{C,2}}(0)$ for some $r > 0$. Let $\mathfrak{D}_\delta$ be a δ-cover for $\mathcal{U}$ in the $\|\cdot\|_{C,2}$-norm, where δ

is chosen so that $4\delta\mathbb{E}_Q\{r+|y|\} \le \epsilon$. We define $\mathfrak{G}_\epsilon : \{((|h_C(A'\mathbf{u}) - y| - \delta)_+)^2\}_{A\in\mathfrak{D}_\delta} \cup \{(|h_C(A'\mathbf{u}) - y| + \delta)^2\}_{A\in\mathfrak{D}_\delta}$.

We proceed to verify (i) and (ii). Let $g = (h_C(A'\mathbf{u}) - y)^2 \in \mathfrak{G}$ be arbitrary. Let $A_0 \in \mathfrak{D}_\delta$ be such that $\|A - A_0\|_{C,2} \le \delta$. Define $\underline{g} = ((|h_C(A_0'\mathbf{u}) - y| - \delta)_+)^2$ and $\overline{g} = (|h_C(A_0'\mathbf{u}) - y| + \delta)^2$. It follows that $\underline{g} \le g \le \overline{g}$, which verifies (i). Next

$$\mathbb{E}\{\overline{g} - \underline{g}\} \le 4\delta\mathbb{E}\{|h_C(A_0'\mathbf{u}) - y|\} \le 4\delta\mathbb{E}\{r + |y|\} \le \epsilon, \tag{B.1}$$

which verifies (ii). □

Proof of Proposition 3.3.7. To simplify notation, $\|\cdot\|$ denotes the operator norm $\|\cdot\|_{C,2}$. By Proposition 3.2.5, the map $A \mapsto \Phi(A, P)$ is differentiable in an open neighborhood around $A^\star$ with derivative $2(h_C(A'\mathbf{u}) - y)\mathbf{u} \otimes \boldsymbol{e}_C(A'\mathbf{u})$. Hence to show that the map is twice differentiable with second derivative Γ, it suffices to show that

$$\lim_{\|D\|\to 0} \frac{1}{\|D\|}\Big\|\mathbb{E}\{2(h_C((A^\star + D)'\mathbf{u}) - y)\mathbf{u} \otimes \boldsymbol{e}_C((A^\star + D)'\mathbf{u})\} - \mathbb{E}\{2(h_C(A^{\star\prime}\mathbf{u}) - y)\mathbf{u} \otimes \boldsymbol{e}_C(A^{\star\prime}\mathbf{u})\} - \Gamma(D)\Big\| = 0.$$

First we note that every component of $\varepsilon(\mathbf{u})\mathbf{u} \otimes \boldsymbol{e}_C((A^\star + D)'\mathbf{u})$ is integrable because $\mathbb{E}[\varepsilon(\mathbf{u})^2] < \infty$, and $\mathbf{u} \otimes \boldsymbol{e}_C((A^\star + D)'\mathbf{u})$ is uniformly bounded. Hence by Fubini's Theorem we have

$$\mathbb{E}\{(h_C(A^{\star\prime}\mathbf{u}) - y)\mathbf{u} \otimes \boldsymbol{e}_C((A^\star + D)'\mathbf{u})\} = \mathbb{E}_{\mathbf{u}}[\mathbb{E}_{\varepsilon(\mathbf{u})}[-\varepsilon(\mathbf{u})\mathbf{u} \otimes \boldsymbol{e}_C((A^\star + D)'\mathbf{u})]] = 0. \tag{B.2}$$

Similarly

$$\mathbb{E}\{(h_C(A^{\star\prime}\mathbf{u}) - y)\mathbf{u} \otimes \boldsymbol{e}_C(A^{\star\prime}\mathbf{u})\} = 0. \tag{B.3}$$

Second by differentiability of the map $A \mapsto \Phi(A, P)$ at $A^\star$ we have

$$\lim_{\|D\|\to 0} \frac{1}{\|D\|}\Big\|\mathbb{E}\left\{(2(h_C((A^\star + D)'\mathbf{u}) - y) - 2(h_C(A^{\star\prime}\mathbf{u}) - y) - 2\langle D, \mathbf{u} \otimes \boldsymbol{e}_C(A^{\star\prime}\mathbf{u})\rangle)\right\}\Big\| = 0.$$

By noting that every component of $\mathbf{u} \otimes \boldsymbol{e}_C((A^\star + D)'\mathbf{u})$ is uniformly bounded, and an application of the Dominated Convergence Theorem, we have

$$\lim_{\|D\|\to 0} \frac{1}{\|D\|}\Big\|\mathbb{E}\Big\{\big(2(h_C((A^\star + D)'\mathbf{u}) - y) - 2(h_C(A^{\star\prime}\mathbf{u}) - y) - 2\langle D, \mathbf{u} \otimes \boldsymbol{e}_C(A^{\star\prime}\mathbf{u})\rangle\big)\mathbf{u} \otimes \boldsymbol{e}_C((A^\star + D)'\mathbf{u})\Big\}\Big\| = 0 \tag{B.4}$$

Third since $h_C(\cdot)$ is continuously differentiable at $A^{\star\prime}\mathbf{u}$ for P-a.e. $\mathbf{u}$, we have $\boldsymbol{e}_C((A^\star + D)'\mathbf{u}) \to \boldsymbol{e}_C(A^{\star\prime}\mathbf{u})$ as $\|D\| \to 0$, for P-a.e. $\mathbf{u}$. By the Dominated Convergence Theorem we have $\mathbb{E}\{\boldsymbol{e}_C((A^\star + D)'\mathbf{u})\} \to \mathbb{E}\{\boldsymbol{e}_C(A^{\star\prime}\mathbf{u})\}$ as $\|D\| \to 0$. It follows that

$$\lim_{\|D\|\to 0} \frac{1}{\|D\|} \Big\| 2\mathbb{E}\left\{\langle \mathbf{u} \otimes \boldsymbol{e}_C(A^{\star\prime}\mathbf{u}), D\rangle \mathbf{u} \otimes \boldsymbol{e}_C((A^\star + D)'\mathbf{u})\right\} - 2\mathbb{E}\left\{\langle \mathbf{u} \otimes \boldsymbol{e}_C(A^{\star\prime}\mathbf{u}), D\rangle \mathbf{u} \otimes \boldsymbol{e}_C(A^{\star\prime}\mathbf{u})\right\} \Big\| = 0. \tag{B.5}$$

The result follows by summing the contributions from (B.4) and (B.5), as well as noting that the expressions in (B.2) and (B.3) vanish. □

Appendix C

PROOFS OF CHAPTER 5

The Appendix is divided as follows. In Section C.1 we describe the relation between the Gaussian distance and the Gaussian width. Next, in Section C.2 we analyze the denoising properties of proximal operators. Finally, in Section C.3 we prove the main results (from Section 5.3) of this paper. As described at the end of Section 5.2.2, we reiterate that the assumption that the set $\mathrm{conv}(\mathscr{A}) \subset \mathbb{R}^q$ contains the origin in its interior holds throughout the Appendix.

C.1 Relationship between Gaussian distance and Gaussian width

The *Gaussian width* of a set $S \subseteq \mathbb{R}^q$ is defined as [66]:

$$\omega(S) := \mathbb{E}_{\mathbf{g}\sim\mathcal{N}(\mathbf{0},I_{q\times q})}\left[\sup_{\mathbf{z}\in S}\langle \mathbf{g}, \mathbf{z}\rangle\right].$$

The next definition that we need in order to relate the Gaussian distance and the Gaussian width is the *skewness* $\kappa_C(\mathbf{x})$ of a norm $\|\cdot\|_C$ at a point $\mathbf{x}$:

$$\kappa_C(\mathbf{x}) := \frac{\|\mathcal{P}_{\partial\|\mathbf{x}\|_C}(\mathbf{0})\|_{\ell_2}}{\|\mathcal{P}_{\text{aff.hull.}(\partial\|\mathbf{x}\|_C)}(\mathbf{0})\|_{\ell_2}},$$

where $\mathcal{P}$ denotes the Euclidean projection and aff.hull. denotes the affine hull. The quantity κ has a natural geometric interpretation: since the subdifferential $\partial\|\mathbf{x}\|_C$ corresponds to a face of the dual norm ball $C^* = \{\mathbf{x} : \|\mathbf{x}\|_C^* \leq 1\}$, the parameter $\kappa_C(\mathbf{x})$ measures the skewness of the face $\partial\|\mathbf{x}\|_C$. It is clear from this interpretation that $\kappa_C(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathbb{R}^q$ whenever the unit ball with respect to the dual norm is suitably symmetric. Examples of such convex sets include the ℓ_1-norm ball, the nuclear-norm ball, the ℓ_∞-norm ball and the spectral-norm ball. Figures C.1 and C.2 illustrate the parameter κ for two different unit-norm balls.

Our final definition relates to yet another convex-geometric concept. The tangent cone $T_C(\mathbf{x})$ at a point $\mathbf{x} \in \mathbb{R}^q$ with respect to the unit ball of the $\|\cdot\|_C$-norm (i.e., the convex set C) when $\|\mathbf{x}\|_C = 1$ is:

$$T_C(\mathbf{x}) := \mathrm{cone}\{\mathbf{z} - \mathbf{x} : \mathbf{z} \in \mathbb{R}^q, \|\mathbf{z}\|_C \leq \|\mathbf{x}\|_C\}.$$

For general unnormalized nonzero points $\mathbf{x} \in \mathbb{R}^q$, the tangent cone with respect to C is $T_C(\mathbf{x}/\|\mathbf{x}\|_C)$.

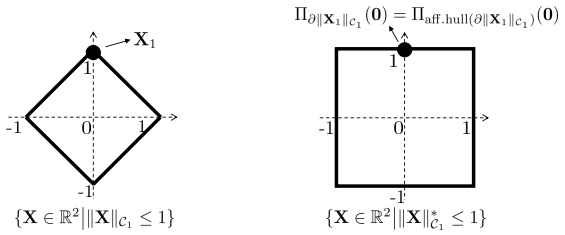


Figure C.1: Figure showing the ℓ_1-norm ball C_1 with parameter $\kappa_{C_1}(\mathbf{x}_1) = 1$.

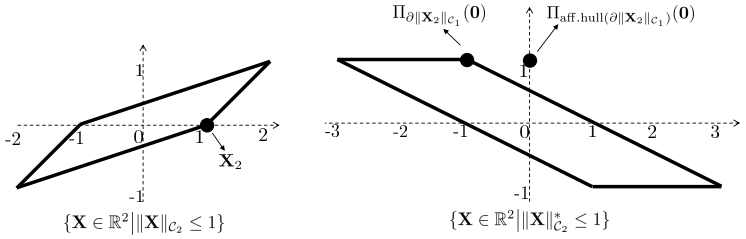


Figure C.2: Figure showing the skewed ℓ_1-norm ball C_2 with parameter $\kappa_{C_2}(\mathbf{x}_2) = \sqrt{2}$.

The next proposition relates the Gaussian distance and the Gaussian width. The result relaxes a "weak decomposability" assumption in [56, Prop.1]. We denote the Euclidean sphere in $\mathbb{R}^q$ by $\mathcal{S}^{q-1}$.

Proposition C.1.1 *The Gaussian distance is bounded above by the Gaussian width as follows:*

$$\eta_C(\mathbf{x}) \leq \omega(T_C(\mathbf{x}) \cap \mathcal{S}^{q-1}) + 3\kappa_C(\mathbf{x}) + 4.$$

Proposition C.1.1 is useful because it relates Theorem 5.3.1 and Proposition 5.3.3 with previously computed bounds on Gaussian widths [35, 56].

The proof of Proposition C.1.1 requires two short lemmas.

Lemma C.1.2 *Suppose* $\mathbf{x} \neq 0$. *Define* $\mathcal{H}$ *to be the affine hull of* $\partial\|\mathbf{x}\|_C$ *and* $\mathbf{w}_0 := \mathcal{P}_{\mathcal{H}}(\mathbf{0})$. *Then* $\langle \mathbf{w} - \mathbf{w}_0, \mathbf{w}_0 \rangle = 0$ *for all* $\mathbf{w} \in \partial\|\mathbf{x}\|_C$ *and* $\|\mathbf{w}_0\|_{\ell_2} > 0$.

Proof. Since $\mathbf{x} \neq \mathbf{0}$, the subdifferential $\partial\|\mathbf{x}\|_C$ is a proper face of the dual norm ball $C^* = \{\mathbf{y} : \|\mathbf{y}\|^*_C \leq 1\}$. Also, since $\mathbf{w}_0 - \mathbf{0}$ is orthogonal to $\mathcal{H}$, we have $\langle \mathbf{w}_0 - \mathbf{0}, \mathbf{w} - \mathbf{w}_0 \rangle = 0$ for all $\mathbf{w} \in \mathcal{H}$. In particular, this holds for all $\mathbf{w} \in \partial\|\mathbf{x}\|_C$. By the assumption that the unit-norm ball has a non-empty interior (see reminder

at the beginning of the Appendix), we have that $\mathbf{0} \in \text{int}(C)$, which implies that $\mathbf{0} \in \text{int}(C^*)$. Consequently, $\mathcal{H}$ does not contain $\mathbf{0}$ and thus $\mathbf{w}_0 \neq \mathbf{0}$. This implies that $\|\mathbf{w}_0\|_{\ell_2} > 0$. □

Lemma C.1.3 *Let $\mathbf{x} \in \mathbb{R}^q$ be an arbitrary nonzero vector. Define $\lambda^\star : \mathbb{R}^q \to \mathbb{R}$ as the function*

$$\begin{aligned}\lambda^\star(\mathbf{g}) &:= \arg\min_{\lambda \geq 0} \text{dist}(\mathbf{g}, \lambda \cdot \partial\|\mathbf{x}\|_C) \\ &= \arg\min_{\lambda \geq 0} \text{dist}^2(\mathbf{g}, \lambda \cdot \partial\|\mathbf{x}\|_C).\end{aligned}$$

Let $\mathcal{H}$ be the affine hull of $\lambda \cdot \partial\|\mathbf{x}\|_C$. Then $\lambda^\star$ is $\frac{1}{\text{dist}(\mathbf{0},\mathcal{H})}$-Lipschitz.

Proof. Let $\mathbf{g}_1, \mathbf{g}_2$ be arbitrary vectors in $\mathbb{R}^q$. Since $\|\mathbf{x}\|_C < \infty$, the subdifferential $\partial\|\mathbf{x}\|_C$ is a closed convex set [120]. Hence we may let $\mathbf{w}_{\mathbf{g}_1}$ be the point in $\partial\|\mathbf{x}\|_C$ such that $\|\mathbf{w}_{\mathbf{g}_1} - \mathbf{g}_1\|_{\ell_2} = \text{dist}(\mathbf{g}_1, \lambda^\star(\mathbf{g}_1) \cdot \partial\|\mathbf{x}\|_C)$. Define $\mathbf{w}_{\mathbf{g}_2}$ similarly. Let $\mathbf{w}_0 = \mathcal{P}_{\mathcal{H}}(\mathbf{0})$ so that

$$\begin{aligned}& \|\lambda^\star(\mathbf{g}_2)\mathbf{w}_{\mathbf{g}_2} - \lambda^\star(\mathbf{g}_1)\mathbf{w}_{g_1}\|_{\ell_2} \\ = \;& \|(\lambda^\star(\mathbf{g}_2) - \lambda^\star(\mathbf{g}_1))\mathbf{w}_0 + (\lambda^\star(\mathbf{g}_2)(\mathbf{w}_{\mathbf{g}_2} - \mathbf{w}_0) + \lambda^\star(\mathbf{g}_1)(\mathbf{w}_0 - \mathbf{w}_{\mathbf{g}_1}))\|_{\ell_2} \\ \geq \;& \langle(\lambda^\star(\mathbf{g}_2) - \lambda^\star(\mathbf{g}_1))\mathbf{w}_0 + (\lambda^\star(\mathbf{g}_2)(\mathbf{w}_{\mathbf{g}_2} - \mathbf{w}_0) + \lambda^\star(\mathbf{g}_1)(\mathbf{w}_0 - \mathbf{w}_{\mathbf{g}_1})), \mathbf{w}_0\rangle \frac{1}{\|\mathbf{w}_0\|_{\ell_2}} \\ = \;& \|\mathbf{w}_0\|_{\ell_2}|\lambda^\star(\mathbf{g}_2) - \lambda^\star(\mathbf{g}_1)|, \qquad \text{(C.1)}\end{aligned}$$

where the last equality follows from Lemma C.1.2. Recall that projection onto a nonempty, closed convex set is nonexpansive, and thus we have $\|\mathbf{g}_2 - \mathbf{g}_1\|_{\ell_2} \geq \|\mathcal{P}_{\cup_{\lambda\geq 0}\{\lambda\cdot\partial\|\mathbf{x}\|_C\}}(\mathbf{g}_2) - \mathcal{P}_{\cup_{\lambda\geq 0}\{\lambda\cdot\partial\|\mathbf{x}\|_C\}}(\mathbf{g}_1)\|_{\ell_2} = \|\mathcal{P}_{\lambda^\star(\mathbf{g}_2)\cdot\partial\|\mathbf{x}\|_C}(\mathbf{g}_2) - \mathcal{P}_{\lambda^\star(\mathbf{g}_1)\cdot\partial\|\mathbf{x}\|_C}(\mathbf{g}_1)\|_{\ell_2} = \|\lambda^\star(\mathbf{g}_2)\mathbf{w}_{\mathbf{g}_2} - \lambda^\star(\mathbf{g}_1)\mathbf{w}_{\mathbf{g}_1}\|_{\ell_2} \geq \|\mathbf{w}_0\|_{\ell_2}|\lambda^\star(\mathbf{g}_2) - \lambda^\star(\mathbf{g}_1)|$. □

Proof of Proposition C.1.1. Our proof is a minor modification of the proof of [56, Prop.1.]. Let $\mathcal{H}$ be the affine hull of $\partial\|\mathbf{x}\|_C$ and $\mathbf{w}_0 = \mathcal{P}_{\mathcal{H}}(\mathbf{0})$. From Lemma C.1.3, we have $\lambda^\star$ is $\frac{1}{\|\mathbf{w}_0\|_{\ell_2}}$-Lipschitz function. Hence by [90, Theorem 5.3], we have $|\lambda^\star(\boldsymbol{\varepsilon}) - \mathbb{E}[\lambda^\star(\bar{\boldsymbol{\varepsilon}})]| \leq c$ for $\bar{\boldsymbol{\varepsilon}} \sim \mathcal{N}(\mathbf{0}, I_{q\times q})$ with probability greater than $1 - 2\exp(-(c\|\mathbf{w}_0\|_{\ell_2})^2/2)$. Suppressing the dependence on $\bar{\boldsymbol{\varepsilon}}$, consider the event $\mathcal{E}_c := \{|\lambda^\star(\boldsymbol{\varepsilon}) - \mathbb{E}[\lambda^\star]| \leq c\}$, and condition on this event. Define $\mathbf{w}_1 := \mathcal{P}_{\partial\|\mathbf{x}\|_C}(\mathbf{0})$ so that $\|\mathbf{w}_1\|_{\ell_2}/\|\mathbf{w}_0\|_{\ell_2} = \kappa(\mathbf{x})$. Let $\mathbf{w}_{\boldsymbol{\varepsilon}} \in \partial\|\mathbf{x}\|_C$ be such that $\|\mathbf{w}_{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}\|_{\ell_2} = \text{dist}(\boldsymbol{\varepsilon}, \lambda^\star(\boldsymbol{\varepsilon}) \cdot \partial\|\mathbf{x}\|_C)$. One has that $\frac{\lambda^\star(\boldsymbol{\varepsilon})}{\mathbb{E}[\lambda^\star]+c}\mathbf{w}_{\boldsymbol{\varepsilon}} + \frac{\mathbb{E}[\lambda^\star]+c-\lambda^\star(\boldsymbol{\varepsilon})}{\mathbb{E}[\lambda^\star]+c}\mathbf{w}_1$ is a convex combination of $\mathbf{w}_1$

and $\mathbf{w}_{\boldsymbol{\varepsilon}}$ (as we condition on $\mathcal{E}_c$), and hence it belongs to $\partial\|\mathbf{x}\|_C$. Then

$$\begin{aligned}
& \operatorname{dist}(\boldsymbol{\varepsilon}, (\mathbb{E}[\lambda^\star] + c) \cdot \partial\|\mathbf{x}\|_C) \\
\overset{\text{(i)}}{\leq} \; & \|\boldsymbol{\varepsilon} - (\lambda^\star(\boldsymbol{\varepsilon})\mathbf{w}_{\boldsymbol{\varepsilon}} + (\mathbb{E}[\lambda^\star] + c - \lambda^\star(\boldsymbol{\varepsilon}))\mathbf{w}_1)\|_{\ell_2} \\
\overset{\text{(ii)}}{\leq} \; & \operatorname{dist}(\boldsymbol{\varepsilon}, \lambda^\star(\boldsymbol{\varepsilon}) \cdot \partial\|\mathbf{x}\|_C)) + \|(\mathbb{E}[\lambda^\star] + c - \lambda^\star(\boldsymbol{\varepsilon}))\mathbf{w}_1\|_{\ell_2} \\
\overset{\text{(iii)}}{\leq} \; & \operatorname{dist}(\boldsymbol{\varepsilon}, \lambda^\star(\boldsymbol{\varepsilon}) \cdot \partial\|\mathbf{x}\|_C)) + 2c\kappa(\mathbf{x})\|\mathbf{w}_0\|_{\ell_2}
\end{aligned} \tag{C.2}$$

where (i) is a consequence of $\frac{\lambda^\star(\boldsymbol{\varepsilon})}{\mathbb{E}[\lambda^\star]+c}\mathbf{w}_{\boldsymbol{\varepsilon}} + \frac{\mathbb{E}[\lambda^\star]+c-\lambda^\star(\boldsymbol{\varepsilon})}{\mathbb{E}[\lambda^\star]+c}\mathbf{w}_1 \in \partial\|\mathbf{x}\|_C$, (ii) follows from the triangle inequality, and (iii) follows from the definition of $\kappa(\mathbf{x})$ and our conditioning on the event $\mathcal{E}_c$. Define the function $m : \mathbb{R}^q \to \mathbb{R}$

$$m(\boldsymbol{\varepsilon}) = \operatorname{dist}(\boldsymbol{\varepsilon}, (\mathbb{E}[\lambda^\star] + c) \cdot \partial\|\mathbf{x}\|_C) - \operatorname{dist}(\boldsymbol{\varepsilon}, \lambda^\star(\boldsymbol{\varepsilon}) \cdot \partial\|\mathbf{x}\|_C).$$

Since $m(\boldsymbol{\varepsilon})$ is the difference of two 1-Lipschitz functions and hence 2-Lipschitz, we have the concentration inequality $\mathbb{P}(m < \mathbb{E}[m] - r) \leq \exp(-r^2/8)$. By setting $r = \sqrt{8\log(1/(1 - 2\exp(-(c\|\mathbf{w}_0\|_{\ell_2})^2/2)))}$ we have $\exp(-r^2/8) = 1-2\exp(-(c\|\mathbf{w}_0\|_{\ell_2})^2/2)$. From (C.2) the event $\{m(\boldsymbol{\varepsilon}) \leq 2c\kappa(\mathbf{x})\|\mathbf{w}_0\|_{\ell_2}\}$ holds with probability greater than $1 - 2\exp(-(c\|\mathbf{w}_0\|_{\ell_2})^2/2)$. Hence it must be the case that

$$\mathbb{E}[m(\boldsymbol{\varepsilon})] \leq 2\kappa(\mathbf{x})c\|\mathbf{w}_0\|_{\ell_2} + \sqrt{8\log(1/(1 - 2\exp(-(c\|\mathbf{w}_0\|_{\ell_2})^2/2)))}. \tag{C.3}$$

Define $N := \cup_{\lambda\geq 0}\{\lambda \cdot \partial\|\mathbf{x}\|_C\}$. We have

$$\begin{aligned}
& \eta_C(\mathbf{x}) \\
= \; & \inf_{\lambda}\left\{\mathbb{E}[\operatorname{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial\|\mathbf{x}\|_C)]\right\} \\
\leq \; & \mathbb{E}[\operatorname{dist}(\boldsymbol{\varepsilon}, (\mathbb{E}[\lambda^\star] + c) \cdot \partial\|\mathbf{x}\|_C)] \\
= \; & \mathbb{E}[\operatorname{dist}(\boldsymbol{\varepsilon}, \lambda^\star(\boldsymbol{\varepsilon}) \cdot \partial\|\mathbf{x}\|_C)] + \mathbb{E}[m(\boldsymbol{\varepsilon})] \\
\overset{\text{(i)}}{=} \; & \mathbb{E}[\operatorname{dist}(\boldsymbol{\varepsilon}, N)] + \mathbb{E}[m(\boldsymbol{\varepsilon})] \\
\overset{\text{(ii)}}{\leq} \; & \mathbb{E}[\operatorname{dist}(\boldsymbol{\varepsilon}, N)] + 2\kappa(\mathbf{x})c\|\mathbf{w}_0\|_{\ell_2} + \sqrt{8\log(1/(1 - 2\exp(-(c\|\mathbf{w}_0\|_{\ell_2})^2/2)))} \\
\overset{\text{(iii)}}{\leq} \; & \{\mathbb{E}[\operatorname{dist}^2(\boldsymbol{\varepsilon}, N)]\}^{1/2} + 2\kappa(\mathbf{x})c\|\mathbf{w}_0\|_{\ell_2} + \sqrt{8\log(1/(1 - 2\exp(-(c\|\mathbf{w}_0\|_{\ell_2})^2/2)))} \\
\overset{\text{(iv)}}{\leq} \; & \omega(T_C(\mathbf{x}) \cap \mathcal{S}^{q-1}) + 1 + 2\kappa(\mathbf{x})c\|\mathbf{w}_0\|_{\ell_2} + \sqrt{8\log(1/(1 - 2\exp(-(c\|\mathbf{w}_0\|_{\ell_2})^2/2)))},
\end{aligned}$$

where (i) follows from the definition of $\lambda^\star$, (ii) follows from (C.3), (iii) follows Jensen's Inequality, and (iv) follows from [5, Proposition 10.1]. We obtain the desired bound by setting $c = 1.5/\|\mathbf{w}_0\|_{\ell_2}$. □

C.2 Analysis of proximal denoising operators

The first result describes a useful monotonicity property of convex functions [100].

Lemma C.2.1 (Monotonicity, [100]) *Let f be a convex function. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^q$. Then for any $\mathbf{z}_i \in \partial f(\mathbf{x}_i), i = 1, 2$, we have*

$$\langle \mathbf{z}_1 - \mathbf{z}_2, \mathbf{x}_1 - \mathbf{x}_2 \rangle \geq 0.$$

Our second result applies this monotonicity property to show that the error of proximal denoising operators is robust to small changes in the underlying signal $\mathbf{x}^\star$. Notice that our proposition also describes the performance of proximal denoisers for *combinations* of structured signals corrupted by noise. This is relevant in our subsequent analysis because the proximal denoiser is applied to averages computed near change-points.

Proposition C.2.2 (Robustness) *Suppose $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}^\star + \boldsymbol{\varepsilon} - \mathbf{x}\|_{\ell_2}^2 + f(\mathbf{x})$ for some convex function f and*

Case 1 (Convex combination of two structured signals): $\mathbf{x}^\star = \mu\mathbf{x}_0^\star + (1-\mu)\mathbf{x}_1^\star$ for some $0 \leq \mu \leq 1$ is a convex combination of two signals $\mathbf{x}_0^\star$ and $\mathbf{x}_1^\star$. Then

$$\mathbb{E}\big[\|\mathbf{x}_0^\star - \hat{\mathbf{x}}\|_{\ell_2}\big] \leq (1-\mu)\|\mathbf{x}_0^\star - \mathbf{x}_1^\star\|_{\ell_2} + \mathbb{E}[\mathrm{dist}(\boldsymbol{\varepsilon}, \partial f(\mathbf{x}_0^\star))].$$

In particular when $\mu = 1$ there is no mixture. In this special case the error bound simplifies to

$$\mathbb{E}\big[\|\mathbf{x}_0^\star - \hat{\mathbf{x}}\|_{\ell_2}\big] \leq \mathbb{E}[\mathrm{dist}(\boldsymbol{\varepsilon}, \partial f(\mathbf{x}_0^\star))].$$

Case 2 (Small perturbation to a structured signal): $\mathbf{x}^\star = \mathbf{x}_0^\star + \boldsymbol{\Delta}$. Then

$$\mathbb{E}\big[\|\mathbf{x}_0^\star - \hat{\mathbf{x}}\|_{\ell_2}\big] \leq \|\boldsymbol{\Delta}\|_{\ell_2} + \mathbb{E}[\mathrm{dist}(\boldsymbol{\varepsilon}, \partial f(\mathbf{x}_0^\star))].$$

Here the expectations are with respect to $\boldsymbol{\varepsilon}$.

Proof. We only prove Case 1 since Case 2 follows from a change of variables. We begin by fixing an $\boldsymbol{\varepsilon}$. From the optimality conditions, we have $\mu\mathbf{x}_0^\star + (1-\mu)\mathbf{x}_1^\star + \boldsymbol{\varepsilon} - \hat{\mathbf{x}} \in \partial\|\hat{\mathbf{x}}\|_C$. Let $\mathbf{z}_0 = \arg\min_{\mathbf{z} \in \partial f(\mathbf{x}_0^\star)} \|\mathbf{z} - \boldsymbol{\varepsilon}\|_{\ell_2}$. From the monotonicity property in Lemma C.2.1 we have

$$\langle \mu\mathbf{x}_0^\star + (1-\mu)\mathbf{x}_1^\star + \boldsymbol{\varepsilon} - \hat{\mathbf{x}} - \mathbf{z}_0, \hat{\mathbf{x}} - \mathbf{x}_0^\star \rangle \geq 0.$$

Rearranging terms and applying the Cauchy-Schwarz inequality, we obtain

$$(1-\mu)\|\mathbf{x}_0^\star - \mathbf{x}_1^\star\|_{\ell_2}\|\mathbf{x}_0^\star - \hat{\mathbf{x}}\|_{\ell_2} + \|\mathbf{z}_0 - \boldsymbol{\varepsilon}\|_{\ell_2}\|\mathbf{x}_0^\star - \hat{\mathbf{x}}\|_{\ell_2} \geq \|\mathbf{x}_0^\star - \hat{\mathbf{x}}\|_{\ell_2}^2.$$

Finally, we divide through by $\|\mathbf{x}_0^\star - \hat{\mathbf{x}}\|_{\ell_2}$ and take expectations on both sides with respect to $\boldsymbol{\varepsilon}$ to obtain the desired result. □

The final result concerns a Lipschitz property of proximal operators. Demonstrating such a property allows us to subsequently appeal to concentration of measure results [90].

Lemma C.2.3 (Proximal operators are non-expansive, Section 5 of [101]) *Suppose f is a convex function. Let $\hat{\mathbf{x}}(\boldsymbol{\varepsilon})$ be the optimal solution of the following optimization problem*

$$\hat{\mathbf{x}}(\boldsymbol{\varepsilon}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}^\star + \boldsymbol{\varepsilon} - \mathbf{x}\|_{\ell_2}^2 + f(\mathbf{x}). \tag{C.4}$$

Then $\|\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2\|_{\ell_2} \geq \|\hat{\mathbf{x}}(\boldsymbol{\varepsilon}_1) - \hat{\mathbf{x}}(\boldsymbol{\varepsilon}_2)\|_{\ell_2}$.

Corollary C.2.4 *Fix an $\mathbf{x}^\star \in \mathbb{R}^q$. Define the function $h : \mathbb{R}^q \to \mathbb{R}$ as $h(\boldsymbol{\varepsilon}) = \|\hat{\mathbf{x}}(\boldsymbol{\varepsilon}) - \mathbf{x}^\star\|_{\ell_2}$, where $\hat{\mathbf{x}}(\boldsymbol{\varepsilon})$ is defined in* (C.4)*. Then the function h is 1-Lipschitz.*

Proof. By applying the triangle inequality twice one has $\|\hat{\mathbf{x}}(\boldsymbol{\varepsilon}_1) - \hat{\mathbf{x}}(\boldsymbol{\varepsilon}_2)\|_{\ell_2} \geq \big|\|\hat{\mathbf{x}}(\boldsymbol{\varepsilon}_1) - \mathbf{x}^\star\|_{\ell_2} - \|\mathbf{x}^\star - \hat{\mathbf{x}}(\boldsymbol{\varepsilon}_2)\|_{\ell_2}\big|$. The result follows from an application of Lemma C.2.3. □

C.3 Proofs of results from Section 5.3

In this section we prove Proposition 5.3.2 (our precursor to Theorem 5.3.1) and Proposition 5.3.3. To simplify notation, we denote $\eta_C(\mathscr{X})$ by η in this section. First we establish a tertiary result that is useful for obtaining a sharper bound on the accuracy of the locations of the estimated change-points.

Proposition C.3.1 *Fix an $\mathbf{x}^\star \in \mathbb{R}^q$. Let $\hat{\mathbf{x}}_0$ and $\hat{\mathbf{x}}_1$ be the optimal solutions to $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}^\star + \boldsymbol{\varepsilon} - \mathbf{x}\|_{\ell_2}^2 + f(\mathbf{x})$ for $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_0$ and $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_1$, respectively. Define the function $j : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$, $j(\boldsymbol{\varepsilon}_0, \boldsymbol{\varepsilon}_1) := \|\hat{\mathbf{x}}_0 - \hat{\mathbf{x}}_1\|_{\ell_2}$. Then j is $\sqrt{2}$-Lipschitz.*

Proof. Let $\{\hat{\mathbf{x}}_0^1, \hat{\mathbf{x}}_1^1\}$ and $\{\hat{\mathbf{x}}_0^2, \hat{\mathbf{x}}_1^2\}$ be the optimal solutions corresponding to the two instantiations $(\boldsymbol{\varepsilon}_0^1, \boldsymbol{\varepsilon}_1^1)$ and $(\boldsymbol{\varepsilon}_0^2, \boldsymbol{\varepsilon}_1^2)$ of the vectors $(\boldsymbol{\varepsilon}_0, \boldsymbol{\varepsilon}_1)$. From Lemma C.2.3, we have $\|\hat{\mathbf{x}}_0^1 - \hat{\mathbf{x}}_0^2\|_{\ell_2} \leq \|\boldsymbol{\varepsilon}_0^1 - \boldsymbol{\varepsilon}_0^2\|_{\ell_2}$ and $\|\hat{\mathbf{x}}_1^1 - \hat{\mathbf{x}}_1^2\|_{\ell_2} \leq \|\boldsymbol{\varepsilon}_1^1 - \boldsymbol{\varepsilon}_1^2\|_{\ell_2}$. By applying the

triangle inequality, we have $\|\hat{\mathbf{x}}_0^1 - \hat{\mathbf{x}}_1^1\|_{\ell_2} \leq \|\hat{\mathbf{x}}_0^1 - \hat{\mathbf{x}}_0^2\|_{\ell_2} + \|\hat{\mathbf{x}}_0^2 - \hat{\mathbf{x}}_1^2\|_{\ell_2} + \|\hat{\mathbf{x}}_1^2 - \hat{\mathbf{x}}_1^1\|_{\ell_2}$. Then

$$\begin{aligned}
|\|\hat{\mathbf{x}}_0^1 - \hat{\mathbf{x}}_1^1\|_{\ell_2} - \|\hat{\mathbf{x}}_0^2 - \hat{\mathbf{x}}_1^2\|_{\ell_2}| &\leq \|\hat{\mathbf{x}}_0^1 - \hat{\mathbf{x}}_0^2\|_{\ell_2} + \|\hat{\mathbf{x}}_1^2 - \hat{\mathbf{x}}_1^1\|_{\ell_2} \\
&\leq \|\boldsymbol{\varepsilon}_0^1 - \boldsymbol{\varepsilon}_0^2\|_{\ell_2} + \|\boldsymbol{\varepsilon}_1^1 - \boldsymbol{\varepsilon}_1^2\|_{\ell_2} \\
&\leq \sqrt{2}\|(\boldsymbol{\varepsilon}_0^1, \boldsymbol{\varepsilon}_1^1) - (\boldsymbol{\varepsilon}_0^2, \boldsymbol{\varepsilon}_1^2)\|_{\ell_2}.
\end{aligned}$$

Hence, j is $\sqrt{2}$-Lipschitz. □

Proof of Proposition 5.3.2. We divide the proof into three parts corresponding to the three events of interest.

Part one [$\mathbb{P}(\mathcal{E}_1^c) \leq 2n^{1-r^2}$]: For each change-point $t \in \tau^\star$, define the following event $\mathcal{E}_{1,t} : \{S_t \geq \gamma\}$. Clearly, $\mathcal{E}_1^c = \bigcup_{t\in\tau^\star} \mathcal{E}_{1,t}^c$. We will prove that $\mathbb{P}(\mathcal{E}_{1,t}^c) \leq 2n^{-r^2}$. By taking a union bound over all $t \in \tau^\star$, we have

$$\mathbb{P}(\mathcal{E}_1^c) = \mathbb{P}(\bigcup_{t\in\tau^\star} \mathcal{E}_{1,t}^c) \leq \sum_{t\in\tau^\star} \mathbb{P}(\mathcal{E}_{1,t}^c) \leq 2|\tau^\star|n^{-r^2} \leq 2n^{1-r^2}.$$

We now prove that $\mathbb{P}(\mathcal{E}_{1,t}^c) \leq 2n^{-r^2}$. Conditioning on the event $\mathcal{E}_{1,t}^c$, and by the triangle inequality, we have

$$\begin{aligned}
\gamma >& \|\hat{\mathbf{x}}[t-\theta+1] - \hat{\mathbf{x}}[t+1]\|_{\ell_2} \\
\geq& - \|\mathbf{x}^\star[t-\theta+1] - \hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2} + \|\mathbf{x}^\star[t+1] - \mathbf{x}^\star[t-\theta+1]\|_{\ell_2} - \|\hat{\mathbf{x}}[t+1] - \mathbf{x}^\star[t+1]\|_{\ell_2}.
\end{aligned}$$

Since $\|\mathbf{x}^\star[t+1] - \mathbf{x}^\star[t-\theta+1]\|_{\ell_2} \geq \Delta_{\min} \geq 2\gamma$, one of the two events $\{\|\hat{\mathbf{x}}[t-\theta+1] - \mathbf{x}^\star[t]\|_{\ell_2} \geq \gamma/2\}$ or $\{\|\hat{\mathbf{x}}[t+1] - \mathbf{x}^\star[t+1]\|_{\ell_2} \geq \gamma/2\}$ must occur. Also, since $t \in \tau^\star$, we have $|t - t'| \geq \theta$ for all $t' \in \tau^\star\backslash\{t\}$. Hence the signal is constant over the time instances $\{t-\theta+1, \ldots, t\}$ and $\{t+1, \ldots, t+\theta\}$. By applying Proposition C.2.2, we have the inequalities $\mathbb{E}[\|\mathbf{x}^\star[t-\theta+1] - \hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2}] \leq \frac{\sigma}{\sqrt{\theta}}\eta$ and $\mathbb{E}[\|\hat{\mathbf{x}}[t+1] - \mathbf{x}^\star[t+1]\|_{\ell_2}] \leq \frac{\sigma}{\sqrt{\theta}}\eta$. Thus

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_{1,t}^c) &\leq \mathbb{P}(\|\hat{\mathbf{x}}[t-\theta+1] - \mathbf{x}^\star[t]\|_{\ell_2} \geq \gamma/2) + \mathbb{P}(\|\hat{\mathbf{x}}[t+1] - \mathbf{x}^\star[t+1]\|_{\ell_2} \geq \gamma/2) \\
&\overset{\text{(i)}}{\leq} \mathbb{P}(\|\hat{\mathbf{x}}[t-\theta+1] - \mathbf{x}^\star[t]\|_{\ell_2} \geq \mathbb{E}[\|\hat{\mathbf{x}}[t-\theta+1] - \mathbf{x}^\star[t]\|_{\ell_2}] + r\sqrt{\sigma^2/\theta}\sqrt{2\log n}) \\
&\quad + \mathbb{P}(\|\hat{\mathbf{x}}[t+1] - \mathbf{x}^\star[t+1]\|_{\ell_2} \geq \mathbb{E}[\|\hat{\mathbf{x}}[t+1] - \mathbf{x}^\star[t+1]\|_{\ell_2}] + r\sqrt{\sigma^2/\theta}\sqrt{2\log n}) \\
&\overset{\text{(ii)}}{\leq} 2\exp(-(r\sqrt{2\log n})^2/2) = 2n^{-r^2}
\end{aligned}$$

where (i) follows from the assumption that $\gamma \geq 2\frac{\sigma}{\sqrt{\theta}}\{\eta_C(\mathscr{X}) + r\sqrt{2\log n}\}$, and (ii) follows from Corollary C.2.4 and from [90, Theorem 5.3].

Part two [$\mathbb{P}(\mathcal{E}_2^c) \leq 2n^{1-r^2}$]: We prove that $\mathbb{P}(\mathcal{E}_2^c) \leq 2n^{1-r^2}$ in essentially the same manner in which we showed that $\mathbb{P}(\mathcal{E}_1^c) \leq 2n^{1-r^2}$. For all $t \in \tau_{\text{far}}$, define $\mathcal{E}_{2,t}$ as the event $\mathcal{E}_{2,t} := \{\|\hat{\mathbf{x}}[t-\theta+1] - \hat{\mathbf{x}}[t+1]\|_{\ell_2} \leq \gamma\}$. Then $\mathcal{E}_2^c = \bigcup_{t\in\tau_{\text{far}}} \mathcal{E}_{2,t}^c$. We will start by proving that $\mathbb{P}(\mathcal{E}_{2,t}^c) \leq 2n^{-r^2}$.

By applying the triangle inequality and conditioning on the event $\mathcal{E}_{2,t}^c$ holding for some $t \in \tau_{\text{far}}$, we have $\|\hat{\mathbf{x}}[t-\theta+1] - \mathbf{x}^\star[t+1]\|_{\ell_2} + \|\mathbf{x}^\star[t+1] - \hat{\mathbf{x}}[t+1]\|_{\ell_2} > \|\hat{\mathbf{x}}[t-\theta+1] - \hat{\mathbf{x}}[t+1]\|_{\ell_2} > \gamma$. Consequently, one of the two events $\{\|\hat{\mathbf{x}}[t-\theta+1] - \mathbf{x}^\star[t+1]\|_{\ell_2} \geq \gamma/2\}$ or $\{\|\mathbf{x}^\star[t+1] - \hat{\mathbf{x}}[t+1]\|_{\ell_2} \geq \gamma/2\}$ must hold. Since $t \in \tau_{\text{far}}$, we have $|t - t^\star| > \theta$ for all $t^\star \in \tau^\star$, and thus the signal is constant over the time instances $\{t-\theta+1,\ldots,t+\theta\}$. By Proposition C.2.2, we have $\mathbb{E}[\|\hat{\mathbf{x}}[t-\theta+1] - \mathbf{x}^\star[t-\theta+1]\|_{\ell_2}] \leq \frac{\sigma}{\sqrt{\theta}}\eta$ and $\mathbb{E}[\|\hat{\mathbf{x}}[t+1] - \mathbf{x}^\star[t+1]\|_{\ell_2}] \leq \frac{\sigma}{\sqrt{\theta}}\eta$. This implies that we have that at least one of the following two events $\{\|\hat{\mathbf{x}}[t-\theta+1]-\mathbf{x}^\star[t-\theta+1]\|_{\ell_2} \geq \mathbb{E}[\|\hat{\mathbf{x}}[t-\theta+1]-\mathbf{x}^\star[t-\theta+1]\|_{\ell_2}] + r\sqrt{\sigma^2/\theta}\sqrt{2\log n}\}$ or $\{\|\hat{\mathbf{x}}[t+1] - \mathbf{x}^\star[t+1]\|_{\ell_2} \geq \mathbb{E}[\|\hat{\mathbf{x}}[t+1] - \mathbf{x}^\star[t+1]\|_{\ell_2}] + r\sqrt{\sigma^2/\theta}\sqrt{2\log n}\}$ holds.

From Corollary C.2.4 and from [90, Theorem 5.3], we have that the probability of either event (corresponding to these two inequalities) occuring is less than $2\exp(-(r\sqrt{2\log n})^2/2) = 2n^{-r^2}$. Thus

$$\mathbb{P}(\mathcal{E}_2^c) = \mathbb{P}\Big(\bigcup_{t\in\tau_{\text{far}}} \mathcal{E}_{2,t}^c\Big) \leq \sum_{t\in\tau_{\text{far}}} \mathbb{P}(\mathcal{E}_{2,t}^c) \leq 2|\tau_{\text{far}}|n^{-r^2} \leq 2n^{1-r^2},$$

as required.

Part three [$\mathbb{P}(\mathcal{E}_3^c) \leq n^{1-r^2}$]: Let us now consider the event $\mathcal{E}_3$. To simplify notation, we define $l := 4r\sqrt{\log n}/\eta$. To prove this part of the proposition, we show a slightly stronger result $\mathbb{P}(\mathcal{E}_3^c) \leq 4\theta|\tau^\star|\exp(-l^2\eta^2/16)$. Since $\theta|\tau^\star| \leq n/4$, our bound would imply that $\mathbb{P}(\mathcal{E}_3^c) \leq n^{1-r^2}$.

For all pairs $(t,\delta) \in \tau_{\text{buffer}}$, define the event $\mathcal{E}_{3,t,\delta} = \big\{\|\hat{\mathbf{x}}[t+1] - \hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2} > \|\hat{\mathbf{x}}[t+1+\delta] - \hat{\mathbf{x}}[t-\theta+1+\delta]\|_{\ell_2}\big\}$. Then $\mathcal{E}_3^c = \bigcup_{(t,\delta)\in\tau_{\text{buffer}}} \mathcal{E}_{3,t,\delta}^c$. We start by proving the following bound

$$\mathbb{P}(\mathcal{E}_{3,t,\delta}^c) \leq 2\exp(-l^2\eta^2/16)$$

for all pairs (t,δ) in τ_{buffer}. Fix one such pair and let Δ_t denote the magnitude of the change at $t \in \tau^\star$. From the triangle inequality and Proposition C.2.2 we have that

$$\begin{aligned}
&\mathbb{E}[\|\hat{\mathbf{x}}[t+1]-\hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2}] \\
&\qquad \geq -\mathbb{E}[\|\hat{\mathbf{x}}[t+1] - \mathbf{x}^\star[t+1]\|_{\ell_2}] \\
&\qquad\quad + \mathbb{E}[\|\mathbf{x}^\star[t+1] - \mathbf{x}^\star[t-\theta+1]\|_{\ell_2}] - \mathbb{E}[\|\mathbf{x}^\star[t-\theta+1] - \hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2}] \\
&\qquad \geq \Delta_t - 2\sqrt{\sigma^2/\theta}\eta.
\end{aligned}$$

Suppose that $\delta \geq 0$. By similarly applying the triangle inequality and Proposition C.2.2 we have

$$\begin{aligned}\mathbb{E}[\|\hat{\mathbf{x}}[t+1+\delta]-\hat{\mathbf{x}}[t-\theta+1+\delta]\|_{\ell_2}] \\ &\leq \mathbb{E}[\|\hat{\mathbf{x}}[t+1+\delta]-\mathbf{x}^\star[t+1]\|_{\ell_2}] + \mathbb{E}[\|\mathbf{x}^\star[t+1]-\hat{\mathbf{x}}[t-\theta+1+\delta]\|_{\ell_2}] \\ &\leq (1-\delta/\theta)\Delta_t + 2\sqrt{\sigma^2/\theta}\eta.\end{aligned}$$

A similar set of computations will show that $\mathbb{E}[\|\hat{\mathbf{x}}[t+1+\delta]-\hat{\mathbf{x}}[t-\theta+1+\delta]\|_{\ell_2}] \leq (1+\delta/\theta)\Delta_t + 2\sqrt{\sigma^2/\theta}\eta$ for $\delta < 0$. Combining these inequalities and using the range of values of δ we have

$$\begin{aligned}&\mathbb{E}[\|\hat{\mathbf{x}}[t+1]-\hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2}] - \mathbb{E}[\|\hat{\mathbf{x}}[t+1+\delta]-\hat{\mathbf{x}}[t-\theta+1+\delta]\|_{\ell_2}] \\ \geq\ & \frac{|\delta|}{\theta}\Delta_t - 4\frac{\sigma}{\sqrt{\theta}}\eta \ \geq\ l\frac{\sigma}{\sqrt{\theta}}\eta. \qquad \text{(C.5)}\end{aligned}$$

Then

$$\begin{aligned}\mathbb{P}(\mathcal{E}^c_{3,t,\delta}) =& \mathbb{P}(\|\hat{\mathbf{x}}[t+1+\delta]-\hat{\mathbf{x}}[t-\theta+1+\delta]\|_{\ell_2} > \|\hat{\mathbf{x}}[t+1]-\hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2}) \\ \overset{\text{(i)}}{\leq}& \mathbb{P}\Big(\|\hat{\mathbf{x}}[t+1+\delta]-\hat{\mathbf{x}}[t-\theta+1+\delta]\|_{\ell_2} - \|\hat{\mathbf{x}}[t+1]-\hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2} \\ &+ \mathbb{E}[\|\hat{\mathbf{x}}[t+1]-\hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2}] - \mathbb{E}[\|\hat{\mathbf{x}}[t+1+\delta]-\hat{\mathbf{x}}[t-\theta+1+\delta]\|_{\ell_2}] \geq \frac{l\sigma}{\sqrt{\theta}}\eta\Big) \\ \overset{\text{(ii)}}{\leq}& \mathbb{P}\left(\mathbb{E}[\|\hat{\mathbf{x}}[t+1]-\hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2}] - \|\hat{\mathbf{x}}[t+1]-\hat{\mathbf{x}}[t-\theta+1]\|_{\ell_2} \geq \frac{l\sigma}{2\sqrt{\theta}}\eta\right) \\ &+ \mathbb{P}\Big(\|\hat{\mathbf{x}}[t+1+\delta]-\hat{\mathbf{x}}[t-\theta+1+\delta]\|_{\ell_2} - \mathbb{E}[\|\hat{\mathbf{x}}[t+1+\delta]-\hat{\mathbf{x}}[t-\theta+1+\delta]\|_{\ell_2}] \\ &\qquad \geq \frac{l\sigma}{2\sqrt{\theta}}\eta\Big) \\ \overset{\text{(iii)}}{\leq}& 2\exp(-l^2\eta^2/16),\end{aligned}$$

where (i) follows from (C.5), (ii) follows from the triangle inequality, and (iii) follows from Proposition C.3.1 and from [90, Theorem 5.3]. Since $\mathcal{E}_3^c = \bigcup_{(t,\delta)\in\tau_{\text{buffer}}} \mathcal{E}^c_{3,t,\delta}$, we have via a union bound

$$\mathbb{P}(\mathcal{E}_3^c) \leq \sum_{(t,\delta)\in\tau_{\text{buffer}}} \mathbb{P}(\mathcal{E}^c_{3,t,\delta}) \leq 2|\tau_{\text{buffer}}|\exp(-l^2\eta^2/16) \leq 4\theta|\tau^\star|\exp(-l^2\eta^2/16).$$

This concludes the proof of Proposition 5.3.2. □

Before proving Proposition 5.3.3 we require a short lemma.

Lemma C.3.2 *Let* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_{q\times q})$. *Then*

$$\operatorname{dist}^2(\boldsymbol{\varepsilon}, \lambda \cdot \partial\|\mathbf{x}\|_C) \leq 2\big(\mathbb{E}\big[\operatorname{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial\|\mathbf{x}\|_C)\big]\big)^2 + 2\sigma^2 t^2$$

with probability greater than $1 - 2\exp(-t^2/2)$.

Proof. The mapping $\boldsymbol{\varepsilon} \mapsto \operatorname{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial\|\mathbf{x}\|_C)$ is nonexpansive and hence 1-Lipschitz. Using Theorem 5.3 from [90], we have

$$\operatorname{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial\|\mathbf{x}\|_C) \leq \mathbb{E}\big[\operatorname{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial\|\mathbf{x}\|_C)\big] + t\sigma \tag{C.6}$$

with probability greater than $1 - \exp(-t^2/2)$. By conditioning on the event corresponding to the inequality (C.6), we apply the arithmetic-geometric-mean inequality and conclude that

$$\operatorname{dist}^2(\boldsymbol{\varepsilon}, \lambda \cdot \partial\|\mathbf{x}\|_C) \leq 2(\mathbb{E}\big[\operatorname{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial\|\mathbf{x}\|_C)\big])^2 + 2t^2\sigma^2$$

with probability greater than $1 - \exp(-t^2/2)$. □

Proof of Proposition 5.3.3. It follows from the proof of Proposition 5.3.2 that the event $\mathcal{E}_1 \cap \mathcal{E}_2$ holds with probability greater than $1 - 4n^{1-r^2}$. Conditioning on the event that $\mathcal{E}_1 \cap \mathcal{E}_2$ holds, the reconstructed signal is constant over the interval $\{t_1 + \theta, \ldots, t_2 - \theta\}$. The result then follows from an application of Lemma C.3.2 and a union bound. □

BIBLIOGRAPHY

[1] A. Agarwal, P. L. Bartlett, and J. C. Duchi. Oracle Inequalities for Computationally Adaptive Model Selection. *CoRR*, abs/1208.0129, 2012.

[2] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli. Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization. *SIAM Journal on Optimization*, 26(4):2775–2799, 2016. doi: 10.1137/140979861.

[3] A. Agarwal, A. Anandkumar, and P. Netrapalli. A Clustering Approach to Learning Sparsely Used Overcomplete Dictionaries. *IEEE Transactions on Information Theory*, 63(1):575–592, 2017. doi: 10.1109/TIT.2016.2614684.

[4] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006. doi: 10.1109/TSP.2006.881199.

[5] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living On The Edge: Phase Transitions in Convex Programs with Random Data. *Information and Inference*, 2014. doi: 10.1093/imaiai/iau005.

[6] A. A. Amini and M. J. Wainwright. High-Dimensional Analysis of Semidefinite Relaxations for Sparse Principal Components. *The Annals of Statistics*, 37(5B):2877–2921, 2009. doi: 10.1214/08-AOS664.

[7] J. Antoch and M. Hušková. *Procedures for the Detection of Multiple Changes in Series of Independent Observations*. Contributions to Statistics. Physica-Verlag HD, 1994. doi: 10.1007/978-3-642-57984-4_1.

[8] S. Arora, R. Ge, and A. Moitra. New Algorithms for Learning Incoherent and Overcomplete Dictionaries. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 35:1–28, 2014.

[9] S. Arora, R. Ge, T. Ma, and A. Moitra. Simple, Efficient, and Neural Algorithms for Sparse Coding. In *Conference on Learning Theory*, 2015.

[10] J. A. D. Aston and C. Kirch. Change Points in High Dimensional Settings. *CoRR*, abs/1409.1771, 2014.

[11] B. Barak, J. A. Kelner, and D. Steurer. Dictionary Learning and Tensor Decomposition via the Sum-of-Squares Method. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*. ACM, 2015. doi: 10.1145/2746539.2746605.

[12] A. R. Barron. Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. doi: 10.1109/18.256500.

[13] M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Applications*. Prentice Halls, 1993.

[14] A. Ben-Tal and A. Nemirovski. On Polyhedral Approximations of the Second-Order Cone. *Mathematics of Operations Research*, 26(2):193–205, 2001. doi: 10.1287/moor.26.2.193.10561.

[15] A. Benveniste and M. Basseville. Detection of Abrupt Changes in Signals and Dynamical Systems: Some statistical aspects. *Lecture Notes in Control and Information Sciences*, 62:143–155, 1984. doi: 10.1007/BFb0004951.

[16] Q. Berthet and P. Rigollet. Optimal Detection of Sparse Principal Components in High Dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013. doi: 10.1214/13-AOS1127.

[17] P. R. Bertrand. A Local Method for Estimating Change Points: the "Hat-Function". *Statistics: A Journal of Theoretical and Applied Statistics*, 34(3): 215–235, 2000. doi: 10.1080/02331880008802714.

[18] P. R. Bertrand, M. Fhima, and A. Guillin. Off-Line Detection of Multiple Change Points by the Filtered Derivative with $p-$Value Method. *Sequential Analysis*, 30:172–207, 2011. doi: 10.1080/07474946.2011.563710.

[19] B. N. Bhaskar, G. Tang, and B. Recht. Atomic Norm Denoising with Applications to Line Spectral Estimation. *CoRR*, abs/1204.0562, 2012.

[20] B. N. Bhaskar, G. Tang, and B. Recht. Atomic Norm Denoising with Applications to Line Spectral Estimation. *IEEE Transactions on Signal Processing*, 61(23):5987–5999, 2013.

[21] P. J. Bickel and E. Levina. Regularized Estimation of Large Covariance Matrices. *The Annals of Statistics*, 36(1):199–227, 2008. doi: 10.1214/009053607000000758.

[22] P. J. Bickel and E. Levina. Covariance Regularization by Thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008. doi: 10.1214/08-AOS600.

[23] A. Birnbaum and S. Shalev-Shwartz. Learning Halfspaces with the Zero-One Loss: Time-Accuracy Tradeoffs. In F. Pereira, C.J.C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 926–934. Curran Associates, Inc., 2012.

[24] T. Blumensath and M. E. Davies. Iterative Hard Thresholding for Compressed Sensing. *Applied and Computational Harmonic Analysis*, 27:265–274, 2009. doi: 10.1016/j.acha.2009.04.002.

[25] E. M. Bronstein. Approximation of Convex Sets by Polytopes. *Journal of Mathematical Sciences*, 153(6):727–762, 2008. doi: 10.1007/s10958-008-9144-x.

[26] A. M. Bruckstein, D. L. Donoho, and M. Elad. From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. *SIAM Review*, 51(1):34–81, 2009. doi: 10.1137/060657704.

[27] T. T. Cai, A. Guntuboyina, and Y. Wei. Adaptive Estimation of Planar Convex Sets. *The Annals of Statistics*, 46(3):1018 – 1049, 2018. doi: doi:10.1214/17-AOS1576.

[28] E. J. Candès and Y. Plan. Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359. doi: 10.1109/TIT.2011.2111771.

[29] E. J. Candès and Y. Plan. Near-Ideal Model Selection by ℓ_1 Minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009. doi: 10.1214/08-AOS653.

[30] E. J. Candès and B. Recht. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. doi: 10.1007/s10208-009-9045-5.

[31] E. J. Candès and T. Tao. Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006. doi: 10.1109/TIT.2006.885507.

[32] E. J. Candès, J. Romberg, and T. Tao. Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006. doi: 10.1109/TIT.2005.862083.

[33] V. Chandrasekaran and M. I. Jordan. Computational and Statistical Tradeoffs via Convex Relaxation. *Proceedings of the National Academy of Sciences*, 110(13):1181–1190, 2013. doi: 10.1073/pnas.1302293110.

[34] V. Chandrasekaran, P. Parillo, and A. S. Willsky. Latent Variable Graphical Model Selection via Convex Optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012. doi: 10.1214/11-AOS949.

[35] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The Convex Geometry of Linear Inverse Problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012. doi: 10.1007/s10208-012-9135-7.

[36] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998. doi: 10.1137/S1064827596304010.

[37] E. Cho. Inner Products of Random Vectors on S^n. *Journal of Pure and Applied Mathematics: Advances and Applications*, 9(1):63–68, 2013.

[38] H. Cho and P. Fryzlewicz. Multiple-Change-Point Detection for High Dimensional Time Series via Sparsified Binary Segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2014. doi: 10.1111/rssb.12079.

[39] Y. S. Chow, H. Robbins, and D. Siegmund. *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, 1971.

[40] M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. In *Advances in Neural Information Processing Systems*, 2013.

[41] L. Danzer. Finite Point-sets On S2 With Minimum Distance As Large As Possible. *Discrete Mathematics*, 60:3 – 66, 1986. doi: doi.org/10.1016/0012-365X(86)90002-6.

[42] S. Dasgupta. The Hardness of k-means Clustering. Technical Report CS2008-0916, University of California, San Diego, 2008.

[43] K. R. Davidson and S. J. Szarek. Local Operator Theory, Random Matrices and Banach Spaces. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook of the Geometry of Banach Spaces*, chapter 8, pages 317–366. Elsevier B. V., 2011.

[44] S. Decatur, O. Goldreich, and D. Ron. Computational Sample Complexity. *SIAM Journal on Computing*, 29:854–879, 1998.

[45] R. A. DeVore and V. N. Temlyakov. Some Remarks on Greedy Algorithms. *Advances in Computational Mathematics*, 5(1):173–187, 1996. doi: 10.1007/BF02124742.

[46] M. Deza and M. Laurent. *Geometry of Cuts and Metrics*. Springer, 1997.

[47] D. L. Donoho. De-noising by Soft-Thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995. doi: 10.1109/18.382009.

[48] D. L. Donoho. Compressed Sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. doi: 10.1109/TIT.2006.871582.

[49] D. L. Donoho. For Most Large Underdetermined Systems of Linear Equations the Minimal ℓ_1-norm Solution Is Also the Sparsest Solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006. doi: 10.1002/cpa.20132.

[50] D. L. Donoho and X. Huo. Uncertainty Principles and Ideal Atomic Decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862.

[51] M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010. doi: 10.1007/978-1-4419-7011-4.

[52] F. Enikeeva and Z. Harchaoui. High-Dimensional Change-Point Detection with Sparse Alternatives. *CoRR*, abs/1312.1900, 2013.

[53] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Department of Electrical Engineering, Stanford University, 2002.

[54] M. Fazel, E. Candès, B. Recht, and P. Parrilo. Compressed Sensing and Robust Recovery of Low Rank Matrices. In *42nd IEEE Asilomar Conference on Signals, Systems and Computers*, 2008.

[55] N. I. Fisher, P. Hall, B. A. Turlach, and G. S. Watson. On the Estimation of a Convex Set From Noisy Data on Its Support Function. *Journal of the American Statistical Association*, 92(437), 1997. doi: 10.2307/2291452.

[56] R. Foygel and L. Mackey. Corrupted Sensing: Novel Guarantees for Separating Structured Signals. *IEEE Transactions on Information Theory*, 60(2): 1223–1247, 2014. doi: 10.1109/TIT.2013.2293654.

[57] P. Fryzlewicz. Wild Binary Segmentation for Multiple Change-Point Detection. *The Annals of Statistics*, 42(6):2243–2281, 2014. doi: 10.1214/14-AOS1245.

[58] P. Gaenssler and W. Stute. Empirical Processes: A Survey of Results for Independent and Identically Distributed Random Variables. *The Annals of Probability*, 7(2):193 – 243, 1979.

[59] Frank Gaillard. Normal Chest CT (Lung Window) – Radiopaedia. https://radiopaedia.org/cases/normal-chest-ct-lung-window.

[60] R. J. Gardner and M. Kiderlen. A New Algorithm for 3D Reconstruction from Support Functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3), 2009.

[61] R. J. Gardner, M. Kiderlen, and P. Milanfar. Convergence of Algorithms for Reconstructing Convex Bodies and Directional Measures. *The Annals of Statistics*, 34(3), 2006. doi: 10.1214/009053606000000335.

[62] A. Garg, L. Gurvits, R. Oliveira, and A. Wigderson. A Deterministic Polynomial Time Algorithm for Non-Commutative Rational Identity Testing with Applications. In *IEEE 57th Annual Symposium on Foundations of Computer Science*, 2016. doi: 10.1109/FOCS.2016.95.

[63] M. Gavish and D. L. Donoho. Optimal Shrinkage of Singular Values. *CoRR*, abs/1405.7511, 2014.

[64] R. Ge, J. D. Lee, and T. Ma. Matrix Completion has No Spurious Local Minimum. In *Advances in Neural Information Processing Systems*, 2016.

[65] D. Goldfarb and S. Ma. Convergence of Fixed-Point Continuation Algorithms for Matrix Rank Minimization. *Foundations of Computational Mathematics*, 11:183–210, 2011. doi: 10.1007/s10208-011-9084-6.

[66] Y. Gordon. On Milman's Inequality and Random Subspaces which Escape Through a Mesh in $\mathbb{R}^n$. In *Geometric Aspects of Functional Analysis*, volume 1317 of *Lecture Notes in Mathematics*, pages 84–106. Springer Berlin Heidelberg, 1988. ISBN 978-3-540-19353-1. doi: 10.1007/BFb0081737.

[67] W. M. Gorman. Estimating Trends in Leontief Matrices. *Unplublished note, referenced in Bacharach (1970)*, 1963.

[68] J. Gouveia and R. R. Thomas. Spectrahedral approximations of convex hulls of algebraic sets. In G. Blekherman, P. A. Parrilo, and R. R. Thomas, editors, *Semidefinite Optimization and Convex Algebraic Geometry*, pages 293–340. MOS-SIAM Series on Optimization, 2013. ISBN 978-1-61197-228-3. doi: 10.1137/1.9781611972290.

[69] J. Gouveia, P. Parrilo, and R. Thomas. Theta Bodies for Polynomial Ideals. *SIAM Journal on Optimization*, 20:2097–2118, 2010. doi: 10.1137/090746525.

[70] J. Gouveia, P. A. Parrilo, and R. Thomas. Lifts of Convex Sets and Cone Factorizations. *Mathematics of Operations Research*, 38(2):248–264, 2013. doi: 10.1287/moor.1120.0575.

[71] J. Gregor and F. R. Rannou. Three-dimensional Support Function Estimation and Application for Projection Magnetic Resonance Imaging. *International Journal of Imaging Systems Technology*, 12:43–50, 2002.

[72] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert. Sample Complexity of Dictionary Learning and Other Matrix Factorizations. *IEEE Transactions on Information Theory*, 61(6):3469–3486, 2015. doi: 10.1109/TIT.2015.2424238.

[73] O. Güler and F. Gürtuna. Symmetry of Convex Sets and its Applications to the Extremal Ellipsoids of Convex Bodies. *Optimization Methods and Software*, 27(4–5):735–759, 2012. doi: 10.1080/10556788.2011.626037.

[74] A. Guntuboyina. Optimal Rates of Convergence for Convex Set Estimation from Support Functions. *The Annals of Statistics*, 40(1):385 – 411, 2012. doi: doi:10.1214/11-AOS959.

[75] L. Gurvits. Classical Complexity and Quantum Entanglement. *Journal of Computer and Systems Sciences*, 69(3):448–484, 2004. doi: 10.1016/j.jcss.2004.06.003.

[76] Z. Harchaoui and C. Lévy-Leduc. Multiple Change-Point Estimation with a Total Variation Penalty. *Journal of the American Statistical Association*, 105 (492):1480–1493, 2010. doi: 10.1198/jasa.2010.tm09181.

[77] J. W. Helton and V. Vinnikov. Linear Matrix Inequality Representation of Sets. *Communications on Pure and Applied Mathematics*, 60(5):654–674, 2007. doi: 10.1002/cpa.20155.

[78] N. J. Higham. Computing the Nearest Correlation Matrix – A Problem from Finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, 2002. doi: 10.1093/imanum/22.3.329.

[79] C. J. Hillar and L.-H. Lim. Most Tensor Problems are NP-Hard. *Journal of the ACM*, 60(6):45:1–45:39, 2013. doi: 10.1145/2512329.

[80] M. Idel. A Review of Matrix Scaling and Sinkhorn's Normal Form for Matrices and Positive Maps. *CoRR*, abs/1609.06349, 2016.

[81] S. Jagabathula and D. Shah. Inferring Rankings Using Constrained Sensing. *IEEE Transactions on Information Theory*, 57(11):7288–7306, 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2165827.

[82] P. Jain, R. Meka, and I. S. Dhillon. Guaranteed Rank Minimization via Singular Value Projection. In *Advances in Neural Information Processing Systems*, 2009.

[83] L. K. Jones. A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training. *The Annals of Statistics*, 20(1):608–613, 1992. doi: 10.1214/aos/1176348546.

[84] T. Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, 1966.

[85] L. Khachiyan and B. Kalantari. Diagonal Matrix Scaling and Linear Programming. *SIAM Journal on Optimization*, 2(4):668–672, 1991. doi: 10.1137/0802034.

[86] M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax Localization of Structural Information in Large Noisy Matrices. In *Neural Information Processing Systems*, 2011.

[87] T. G. Kolda and B. W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, 2009. doi: 10.1137/07070111X.

[88] M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008.

[89] J. B. Lasserre. Global Optimization with Polynomials and the Problem of Moments. *SIAM Journal on Optimization*, 11:796–817, 2001. doi: 10.1137/S1052623400366802.

[90] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical surveys and monographs*. American Mathematical Society, 2001. ISBN 9780821837924.

[91] A. S. Lele, S. R. Kulkani, and A. S. Willsky. Convex-polygon Estimation from Support-line Measurements and Applications to Target Reconstruction from Laser-radar Data. *Journal of the Optical Society of America, Series A*, 9:1693–1714, 1992.

[92] N. Linial, A. Samorodnitsky, and A. Wigderson. A Deterministic Strongly Polynomial Algorithm for Matrix Scaling and Approximate Permanents. *Combinatorica*, 20(4):545–568, 2000. doi: 10.1007/s004930070007.

[93] S. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129 – 137, 1982.

[94] G. Lorden. Procedures for Reacting to a Change in Distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971. doi: 10.1214/aoms/1177693055.

[95] M. Mahajana, P. Nimbhorkara, and K. Varadarajan. The Planar k-means Problem is NP-hard. *Theoretical Computer Science*, 442, 2012. doi: 10.1016/j.tcs.2010.05.034.

[96] J. Mairal, F. Bach, and J. Ponce. Sparse Modeling for Image and Vision Processing. *Foundations and Trends in Computer Graphics and Vision*, 8 (2–3):85–283, 2014. doi: 10.1561/0600000058.

[97] O. L. Mangasarian and B. Recht. Probability of Unique Integer Solution to a System of Linear Equations. *European Journal of Operational Research*, 214(1):27 – 30, 2011. doi: http://dx.doi.org/10.1016/j.ejor.2011.04.010.

[98] M. Marcus and B. N. Moyls. Transformations on Tensor Product Spaces. *Pacific Journal of Mathematics*, 9(4):1215–1221, 1959.

[99] N. Meinhausen and P. Bühlmann. High-Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006. doi: 10.1214/009053606000000281.

[100] G. Minty. On the Monotonicity of the Gradient of a Convex Function. *Pacific Journal of Mathematics*, 14(1):243–247, 1964.

[101] J. J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

[102] B. K. Natarajan. Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, 24(2):227–234, 1993. doi: 10.1137/S0097539792240406.

[103] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM Studies in Applied and Numerical Mathematics, 1994. doi: 10.1137/1.9781611970791.

[104] B. A. Olshausen and D. J. Field. Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature*, 381: 607–609, 1996. doi: 10.1038/381607a0.

[105] S. Oymak and B. Hassibi. Tight Recovery Thresholds and Robustness Analysis for Nuclear Norm Minimization. In *IEEE International Symposium on Information Theory*, pages 2323 – 2327, 2011. doi: 10.1109/ISIT.2011.6033977.

[106] S. Oymak and B. Hassibi. On a Relation between the Minimax Risk and the Phase Transitions of Compressed Recovery. In *50th Annual Allerton Conference on Communication, Control, and Computing*, pages 1018–1025, 2012. doi: 10.1109/Allerton.2012.6483330.

[107] S. Oymak and B. Hassibi. Sharp MSE Bounds for Proximal Denoising. *Foundations of Computational Mathematics*, 16(4):965–1029, 2016. doi: 10.1007/s10208-015-9278-4.

[108] E. A. Page. Continuous Inspection Schemes. *Biometrika*, 41:100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

[109] N. Parikh and S. Boyd. Proximal Algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014. doi: 10.1561/2400000003.

[110] P. A. Parrilo. *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*. PhD thesis, California Institute of Technology, 2000.

[111] G. Pisier. Remarques sur un résultat non publié de B. Maurey. *Séminaire Analyse fonctionnelle (dit "Maurey-Schwartz")*, pages 1–12, 1981.

[112] D. Pollard. Strong Consistency of k-Means Clustering. *The Annals of Statistics*, 9(1):135–140, 1981.

[113] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[114] H. V. Poor and O. Hadjiliadis. *Quickest Detection*. Cambridge University Press, 2008.

[115] J. L. Prince and A. S. Willsky. Reconstructing Convex Sets from Support Line Measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:377–389, 1990.

[116] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(3):471–501, 2010. doi: 10.1137/070697835.

[117] B. Recht, W. Xu, and B. Hassibi. Null Space Conditions and Thresholds for Rank Minimization. *Mathematical Programming*, 127(1):175–202, 2011. doi: 10.1007/s10107-010-0422-2.

[118] J. Renegar. *A Mathematical View of Interior-Point Methods in Convex Optimization*. MOS-SIAM Series on Optimization, 2001. doi: 10.1137/1.9780898718812.

[119] J. Renegar. Hyperbolic Programs and their Derivative Relaxations. *Foundations of Computational Mathematics*, 6:59–79, 2006. doi: 10.1007/s10208-004-0136-z.

[120] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[121] M. Rudelson and R. Vershynin. Sparse Reconstruction by Convex Relaxation: Fourier and Gaussian Measurements. In *40th Annual Conference on Information Sciences and Systems*, pages 207–212, 2006. doi: 10.1109/CISS.2006.286463.

[122] K. Schnass. On the Identifiability of Overcomplete Dictionaries via the Minimisation Principle Underlying K-SVD. *Applied and Computational Harmonic Analysis*, 37(3):464–491, 2014. doi: 10.1016/j.acha.2014.01.005.

[123] K. Schnass. Convergence Radius and Sample Complexity of ITKM Algorithms for Dictionary Learning. *Applied and Computational Harmonic Analysis*, 2016. doi: 10.1016/j.acha.2016.08.002.

[124] R. Schneider. *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge University Press, 1993.

[125] K. Schütte and B. L. van der Waerden. Auf welcher Kugel haben 5, 6, 7, 8 oder 9 Punkte mit Mindestabstand Eins Platz? *Mathematische Annalen*, 123: 96 – 124, 1951. doi: doi.org/10.1007/BF02054944.

[126] R. Servedio. Computational Sample Complexity and Attribute-Efficient Learning. *Journal of Computer and Systems Sciences*, 60:161–178, 2000. doi: 10.1006/jcss.1999.1666.

[127] P. Shah, B. N. Bhaskar, G. Tang, and B. Recht. Linear System Identification via Atomic Norm Regularization. In *51st IEEE Conference on Decisions and Control*, 2012.

[128] S. Shalev-Shwartz, O. Shamir, and E. Tromer. Using More Data to Speed Up Training Time. In *Conference on Artificial Intelligence and Statistics*, 2012.

[129] D. Shender and J. Lafferty. Computation-Risk Tradeoffs for Covariance-Thresholded Regression. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 28(3):756–764, 2013.

[130] H. D. Sherali and W. P. Adams. A Hierarchy of Relaxations between the Continuous and Convex Hull Representations for Zero-One Programming Problems. *SIAM Journal on Discrete Mathematics*, 3:411–430, 1990. doi: 10.1137/0403036.

[131] A. N. Shiryaev. On Optimum Methods in Quickest Detection Problems. *Theory of Probability and its Applications*, 8(1):22–46, 1963. doi: 10.1137/1108002.

[132] R. Sinkhorn. A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964. doi: 10.1214/aoms/1177703591.

[133] S. Smale. Mathematical Problems for the Next Century. *The Mathematical Intelligencer*, 20(2):7 – 15, 1998.

[134] D. A. Spielman, H. Wang, and J. Wright. Exact Recovery of Sparsely-Used Dictionaries. *Journal on Machine Learning and Research: Workshop and Conference Proceedings*, 23(37):1–18, 2012.

[135] H. Stark and H. Peng. Shape Estimation in Computer Tomography from Minimal Data. *Journal of the Optical Society of America, Series A*, 5(3): 331–343, 1988.

[136] G. Stengle and J. E. Yukich. Some New Vapnik-Chervonenkis classes. *The Annals of Statistics*, 17(4):1441 – 1446, 1989.

[137] G. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[138] M. Stojnic. Various Thresholds for ℓ_1-Optimization in Compressed Sensing. *CoRR*, abs/0907.3666, 2009.

[139] J. Sun, Q. Qu, and J. Wright. Complete Dictionary Recovery over the Sphere I: Overview and the Geometric Picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017. doi: 10.1109/TIT.2016.2632162.

[140] J. Sun, Q. Qu, and J. Wright. Complete Dictionary Recovery over the Sphere II: Recovery by Riemannian Trust-region Method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2017. doi: 10.1109/TIT.2016.2632149.

[141] J. Sun, Q. Qu, and J. Wright. A Geometric Analysis of Phase Retrieval. *Foundations of Computational Mathematics*, 2017. doi: 10.1007/s10208-017-9365-9.

[142] P. M. L. Tammes. On the Origin of Number and Arrangement of the Places of Exit on the Surface of Pollen-Grains. *Recueil des travaux botaniques néerlandais*, 27:1 – 87, 1930.

[143] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed Sensing Off the Grid. *IEEE Transactions on Information Theory*, 59(11):7465–7490, 2013. ISSN 0018-9448. doi: 10.1109/TIT.2013.2277451.

[144] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[145] K. C. Toh, M. J. Todd, and R. H. Tütüncü. SDPT3 – a MATLAB Software Package for Semidefinite Programming. *Optimization Methods and Software*, 11:545–581, 1999. doi: 10.1080/10556789908805762.

[146] J. A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

[147] L. Tunçel. Potential Reduction and Primal-Dual Methods. In H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors, *Handbook of Semidefinite Programming – Theory, Algorithms, and Applications*, chapter 9. Kluwer's International Series in Operations Research and Management Science, 2000. doi: 10.1007/978-1-4615-4381-7.

[148] D. Vainsencher, S. Mannor, and A. M. Bruckstein. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12, 2011.

[149] V. V. Veeravalli and T. Banerjee. Quickest Change Detection. *CoRR*, abs/1210.5552, 2012.

[150] M. J. Wainwright. Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using L1-constrained Quadratic Programming (Lasso). *IEEE Transactions on Information Theory*, 55(5), 2009. doi: 10.1109/TIT.2009.2016018.

[151] Y. Xie, J. Huang, and R. Willett. Change-Point Detection for High-Dimensional Time Series With Missing Data. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):12–27, 2013. doi: 10.1109/JSTSP.2012.2234082.

[152] M. Yannakakis. Expressing Combinatorial Optimization Problems by Linear Programs. *Journal of Computer and System Sciences*, 43:441–466, 1991. doi: 10.1016/0022-0000(91)90024-Y.