The Dissertation Committee for Maria Soledad Villar Lozano
certifies that this is the approved version of the following dissertation:

Relax, descend, and certify: optimization techniques for typically tractable data problems

Committee:

Rachel Ward, Supervisor

Afonso S. Bandeira

Andrew J. Blumberg

Arie Israel

Relax, descend, and certify: optimization techniques for typically tractable data problems

by

Maria Soledad Villar Lozano, B.S., B.Eng., M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2017

En memoria de mi abuela Nelly.

Acknowledgments

I have the deepest feelings of gratitude and admiration to my advisor, Rachel Ward, who has taught me, encouraged me, shared great math ideas, and given me amazing opportunities throughout my PhD.

I am also very grateful to have such talented mentors like Dustin Mixon and Afonso Bandeira. I truly appreciate their support, our very enlightening math discussions, and their friendship.

I am very fortunate to have amazing collaborators. This thesis is in fact based on different projects with multiple coauthors: Pranjal Awasthi, Afonso Bandeira, Andrew Blumberg, Timothy Carson, Moses Charikar, Ravishankar Krishnaswamy, Takayuki Iguchi, Dustin Mixon, Jesse Peterson, and of course Rachel Ward. I am thankful I had the opportunity to work with every single one of them.

I thank my thesis' committee members for the feedback provided. In particular I am grateful to Arie Israel for his sharp comments that greatly improved this manuscript.

I appreciate the help and support from faculty and staff from Univerisity of Texas at Austin, in particular Andrew Blumberg, Bubacarr Bah, Dan Knopf, Sandra Cattlet, Thomas Chen and Elisa Bass.

I also want to thank my undergrad and master's advisor, Gonzalo

Tornaría, for his central role in my mathematical education and in me coming to Austin.

The Uruguay Math Olympiads has played an important part inspiring my mathematical curiosity from a very young age. I want to thank Ariel Affonso and everyone that made this unusual career path even possible for me, and everyone who through the Uruguay Math Olympiads is inspiring young kids to love math as I do.

I thank my friends in Uruguay and Austin for their invaluable friendship, and in particular Tim for making me happier every day.

Lastly but most importantly I want to acknowledge my family, specially my parents Pilar and José, and my sisters Andrea and Florencia, for being the most important invariant in my life.

Relax, descend, and certify: optimization techniques for typically tractable data problems

Publication No. _____

Maria Soledad Villar Lozano, Ph.D.
The University of Texas at Austin, 2017

Supervisor: Rachel Ward

In this thesis we explore different mathematical techniques for extracting information from data. In particular we focus in machine learning problems such as clustering and data cloud alignment. Both problems are intractable in the "worst case", but we show that convex relaxations can efficiently find the exact or almost exact solution for classes of "typical" instances. We study different roles that optimization techniques can play in understanding and processing data. These include efficient algorithms with mathematical guarantees, a posteriori methods for quality evaluation of solutions, and algorithmic relaxation of mathematical models. We develop probabilistic and data-driven techniques to model data and evaluate performance of algorithms for data problems.

Table of Contents

List of Figures

Chapter 1

Introduction

The problem of extracting knowledge from data is very relevant these days. The classical statistical approach for this kind of problem consists in the following steps (i) acquiring and processing data, (ii) formulating a statistical model depending on a few parameters, (iii) formulating a likelihood function of the parameters given the data, (iv) solving the optimization problem (finding the best parameters that maximize the likelihood for the given data).

The machine learning approach takes many of its techniques and ideas from statistics, but it formulates the problems in a slightly different way. For instance, supervised machine learning deals with large amounts of labeled data $\{(d_i, l_i)\}_{i=1}^{n}$. Here $d_i \in \mathcal{D}$ corresponds to data (for example an image), and $l_i \in \mathcal{L}$ represents its label (for example 'dog'). Supervised machine learning typically has a training step, with the purpose of inferring the *best* function $f : \mathcal{D} \to \mathcal{L}$ that adjusts to known information $(d_i, f(d_i) = l_i)$. The function f is used to predict labels for new data. In a different manner, unsupervised machine learning finds hidden structures or patterns in unlabeled data, like clusters or manifolds.

Regardless of the approach to data modeling and processing, all these problems, at the end, amount to solving an optimization problem that can be expressed in the form

$$\begin{aligned} &\text{minimize} && f_D(x) \\ &\text{subject to} && x \in \mathcal{S}. \end{aligned} \tag{1.1}$$

where $D \in \mathcal{D}$ represents the data of our problem in the universe of all possible data $\mathcal{D}$, and x represents a potential answer to the question we are asking about the data. The set of all possible answers is $\mathcal{S}$ which can be of many different shapes.

Optimization problems arising from data can be intractable in many cases. However, the computational complexity of a problem measures the amount of time or space that it takes to solve its *hardest instance*. For the problems we study in this thesis (and many other problems existing in the literature) it turns out that even NP-hard problems can be solved in polynomial-time for a large number of instances $D \in \mathcal{D}$. Compressed sensing is a famous example of this phenomenon [21].

In this thesis we focus on two data problems: clustering and point cloud matching. We study different approaches to their underlying optimization problem that allow us to:

(i) Find the exact solution of (1.1) for a non-trivial subset of $\mathcal{D}$.

(ii) Find an approximate solution of (1.1) for a larger subset of $\mathcal{D}$, with explicit approximation bounds.

(iii) Leverage fast heuristic algorithms and mathematical proofs to develop quasi-linear time algorithms (in D) that provide the exact solution of (1.1) for a non-trivial subset of $\mathcal{D}$.

(iv) Substitute NP-hard functions by tractable proxies that preserve many interesting properties of the original functions.

This description may seem pretty abstract for now. A more concrete explanation is provided in Section 1.4, where we summarize the ideas for the specific problems. However, these techniques may be applied to seemingly any data-related problem.

1.1 Clustering

Clustering is a central problem in unsupervised machine learning. It consists of partitioning a given set P, a finite set of points of a metric space (X, d), into k subsets such that some dissimilarity function is minimized. The dissimilarity function is in general chosen with an application in mind. Due to the nature of most machine learning problems, identifying similar data is a main learning step and therefore many complex algorithms rely on clustering subroutines.

The clustering objective known as k-means is one of the most common for data in Euclidean space, and k-medians is widely used in general metric spaces (like tree spaces). Figure 1.1 depicts both problems.

k-means In the euclidean k-*means* problem, the set of points P is in $\mathbb{R}^m$ and

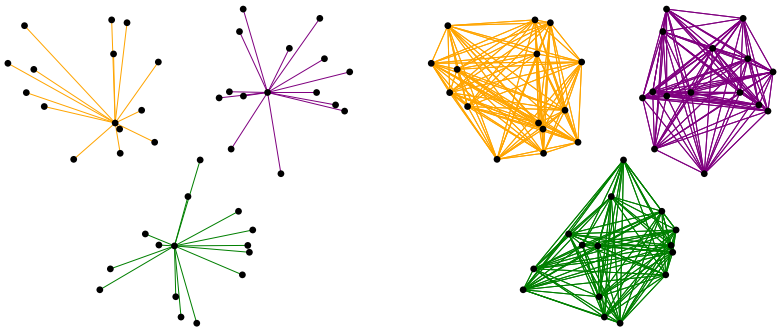


Figure 1.1: Illustration of k-medians and k-means
The k-medians objective (left) minimizes the sum of distances from points to their representative data points. The k-means objective (right) minimizes the average of the squared euclidean distances of all points within a cluster.

the distance is the Euclidean distance $d(x_i, x_j) = \|x_i - x_j\|$. The goal is to partition the finite set $P = \{x_1, \ldots, x_N\}$ in k clusters such that the sum of the squared euclidean distances to the average point of each cluster (not necessarily a point in P) is minimized. Let $A_1, A_2, \ldots, A_k$ denote a partitioning of the the indices $[N] = \{1, \ldots, N\}$ into k subsets; if $c_t = \frac{1}{|A_t|} \sum_{j \in A_t} x_j$ denotes the centroid of the cluster t, then the k-means problem reads

$$\underset{A_1 \cup \cdots \cup A_k = [N]}{\text{minimize}} \sum_{t=1}^{k} \sum_{i \in A_t} \|x_i - c_t\|^2$$

By expanding the square one obtains the identity $\sum_{i \in A_t} \|x_i - c_t\|^2 = \frac{1}{2} \frac{1}{|A_t|} \sum_{i,j \in A_t} \|x_i - x_j\|^2$, which allows us to re-express the k-means prob-

lem as the following optimization problem:

$$\underset{A_1 \cup \cdots \cup A_k = [N]}{\text{minimize}} \ \frac{1}{2} \sum_{t=1}^{k} \frac{1}{|A_t|} \sum_{i,j \in A_t} \|x_i - x_j\|^2. \qquad \text{(k-means)}$$

k-medians The k-*medians* (also known as k-medoids) objective is defined for general metric spaces, where the notion of centroid may not exist. In this setting, clusters are specified by *centers*: k representative points *from within the set* P denoted by $c_1, c_2, \ldots, c_k$. The corresponding partitioning is obtained by assigning each point to its closest center. The cost incurred by a point is the distance to its assigned center, and the goal is to find k center points that minimize the sum of the costs of the points in P:

$$\underset{\{c_1, c_2, \ldots, c_k\} \subset P}{\text{minimize}} \ \sum_{i=1}^{n} \min_{t=1,\ldots,k} d(x_i, c_t) \qquad \text{(k-medians)}$$

Both problems can be expressed as an optimization problem of the form (1.1) where $D = P$ and $\mathcal{S}$ is a discrete set in correspondence with all possible partitions of the points in k clusters. Unfortunately, the combinatoric nature of the set of all possible partitions result in both problems being NP-hard [6, 36]. However, being NP-hard is a statement about the hardest instance of the problem; and there is a line of work that claims that clustering is not hard when data is naturally clustered [12].

In fact, the geometric nature of the k-means problem allows a simple and widely used alternating minimization algorithm known as Lloyd's algorithm [43]. Lloyd's algorithm consists of the following steps:

1. select random data points as centers,
2. assign each point to the closest center,
3. recompute centers,

where steps 2 and 3 are repeated until convergence. Lloyd's algorithm is fast but it often converges to a suboptimal clustering, a local minimizer of (k-means). Not only that, but the output of the algorithm provides no information of how far from optimal it may be.

1.1.1 A remark on finding the number of clusters

Both the k-means and k-medians problem formulations assume that the number of clusters k is a known parameter. Sometimes k is given by the problem, for example in the hand-written digits data set that we study in Section 6.5.1, the number of clusters is 10. However, in many problems the number of clusters is not known a priori and should be estimated.

There exists a few techniques that allow us to find the number of clusters. One of the first methods to estimate the number of clusters is the *elbow method*, that can be traced back to [60]. Informally speaking, the method consists of computing the k-means value for different values of k and essentially choosing $k^\star$ to be such that one does not gain too much when setting the number of clusters $k = k^\star + 1$ and does not lose too much when with $k = k^\star - 1$.

Many methods have been developed since then. For example [69] presents a semidefinite program for clustering that chooses the number of

clusters, and [42] presents a spectral method to find the number of clusters.

1.2 Gromov-Hausdorff distance and point cloud matching

In order to study the convergence of sequences of metric spaces, Gromov introduced what is now called the Gromov-Hausdorff metric [30]. Roughly speaking, this metric generalizes the classic Hausdorff distance between a pair of subsets of an ambient metric space, to a distance between a pair of arbitrary metric spaces. This is done by embedding these metric spaces into a third space and taking an infimum over all such embeddings. The Gromov-Hausdorff metric has been of theoretical importance in geometric group theory and is at the heart of the subject of "metric geometry".

More recently, the Gromov-Hausdorff distance has been proposed as a basic method for comparing *point clouds* [48]. A point cloud is simply a finite metric space (often presented as a subset of $\mathbb{R}^m$); this is a fundamental and ubiquitous representation of data. Geometric examples, where the point cloud represents samples from some smooth geometric object, arise from various kinds of shape acquisition devices. Examples with less obvious intrinsic geometric structure are frequently generated by biological data (e.g., collections of gene expression vectors). Given two point clouds, a natural question is to determine if they are related by some isometric transformation; if not, one might wish to know a quantitative measure of their difference.

Another version of this sort of problem is known as the point registration problem (also sometimes referred to as point matching and network alignment). Point registration consists in finding a correspondence between point sets or graphs such that a certain cost function is minimized. It appears in computer vision problems like shape matching [25], computational biology [22], and general pattern recognition problems. In some applications, registering or aligning is particularly challenging since there is no explicit correspondence between the sets, often because deformation has occurred or they have different numbers of points. In such cases it is natural to consider a metric on point clouds that is defined in terms of correspondences between point clouds together; the Gromov-Hausdorff distance can be described in terms of a minimax expression over correspondences between the metric spaces, and so is potentially suitable for this purpose.

Unfortunately, exact computation of the Gromov-Hausdorff distance is essentially intractable; it involves the solution of an NP-hard optimization problem. As a consequence, it is natural to consider relaxations. In [47], Mémoli studied a relaxation referred to as the Gromov-Wasserstein distance — this distance is closely related to distances motivated by optimal transport problems [44, 59], to a "distance distribution" metric defined by Gromov, and also to the cut distance of graphons. Unfortunately, computing the Gromov-Wasserstein distance still requires solving a non-convex optimization problem which does not appear to have attractive performance characteristics in practice.

1.3 Relax-and-round versus exact recovery

A widespread idea to tackle these combinatoric optimization problems is known as the *relax and round* paradigm. It consists in augmenting the discrete domain $\mathcal{S}$ to a larger set $\bar{\mathcal{S}}$ where one can use optimization algorithms; solve the optimization problem in the larger set, obtaining $\bar{x} \in \bar{\mathcal{S}}$; and then round the solution of the relaxed problem into a feasible point of the original problem $\bar{x} \mapsto x^\star \in \mathcal{S}$.

If the function f_D is convex, relaxing the optimization problem into a convex set $\bar{\mathcal{S}}$ results in a problem with a unique local minimizer, where interior point methods [51] are guaranteed to converge to global optima of the relaxed problem. If we provide an explicit bound for the difference $\|\bar{x} - x^\star\|$ and a Lipschitz constant for f_D, we obtain an approximation algorithm for (1.1) with explicit error bounds [63].

Sometimes we may also want to relax $\mathcal{S}$ to a non-convex set, for instance, a smooth manifold, where algorithms can be implemented very efficiently [14] but a priori are only guaranteed to converge to local optima (though under some hypothesis manifold optimization algorithms had been proven to converge to global optimizers [15]).

If the solution $\bar{x}$ of relaxed optimization problem in $\bar{\mathcal{S}}$ happens to be feasible for the original problem (1.1) (i.e.: $\bar{x} \in \mathcal{S}$), then $\bar{x}$ is also optimal for (1.1). In such case we say the relaxation is *tight* or that the relaxation has an *integral* solution.

1.4 Main contributions

1.4.1 Relaxations of the k-means problem

As mentioned before, one can rewrite the k-means problem as

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\operatorname{Tr}(DX) \\ \text{subject to} \quad & X := \sum_{t=1}^{k} \frac{1}{|A_t|} 1_{A_t} 1_{A_t}^\top, \end{aligned} \tag{1.2}$$

where D is an $n \times n$ matrix such that $D_{ij} = \|x_i - x_j\|^2$, and X is a projection matrix into the span of the indicator vectors of each cluster (i.e.: $(1_{A_t})_i = 1$ if $x_i \in A_t$ and 0 otherwise).

An equivalent formulation for k-means is the following optimization in the set of rank k matrices:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\operatorname{Tr}(DYY^\top) \\ \text{subject to} \quad & Y \in \mathbb{R}^{n\times k},\ YY^\top 1 = 1, \\ & Y^\top Y = I_k,\ Y \geqslant 0. \end{aligned} \tag{1.3}$$

Here the constraint $Y^\top Y = I_k$ means that Y has orthonormal columns. Using that $Y \geqslant 0$ entry wise, we obtain that $Y_{ij} \neq 0$ implies that $Y_{ik} = 0$ for all $k \neq j$, so Y has exactly one nonnegative entry per row. The constraint $YY^\top 1 = 1$ implies that the vector $1 \in \mathbb{R}^n$ belongs to the span of the columns of Y. Therefore if $Y_{ij} \neq 0$ and $Y_{lj} \neq 0$ then $Y_{ij} = Y_{lj}$. This shows that if Y is feasible for (1.3) then $X = YY^\top$ is feasible for (1.2).

Optimization problems (1.2) and (1.3) are equivalent to k-means, which is an NP-hard problem [6]. A typical way to tackle such hard problems is to

relax the discrete feasible set to a larger set, then use analytic tools to solve the larger problem, and finally round a solution of the larger problem into a feasible solution for the original problem.

For instance, the *spectral clustering* technique is based on the following relaxation of (1.3):

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\operatorname{Tr}(DYY^\top) \\ &\text{subject to} && Y \in \mathbb{R}^{n\times k},\ Y^\top Y = I_k. \end{aligned} \tag{1.4}$$

Note that the solution of (1.4) is a matrix with columns consisting of the top k eigenvectors of D.

In general, spectral clustering algorithms replace the matrix D by a matrix $-K$, where K corresponds to the Gram matrix of the points mapped to a higher dimensional space (i.e.: $K_{ij} = \langle \phi(x_i), \phi(x_j)\rangle$ for $\phi : \mathbb{R}^n \to \mathbb{R}^N$.) One particularly common implementation uses the Gaussian kernel: $K_{ij} = \exp(-\|x_i - x_j\|^2/\sigma^2)$.

Another relaxation of k-means, that we study in depth in this thesis, is Peng and Wei's k-means SDP [56], which solves

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\operatorname{Tr}(DX) \\ &\text{subject to} && \operatorname{Tr} X = k,\ X\mathbf{1} = \mathbf{1},\ X \geqslant 0,\ X \succeq 0, \end{aligned} \tag{1.5}$$

where $X \succeq 0$ means that X is symmetric and positive semidefinite. Note that the results from [23] indicate that the constraint $X \geqslant 0$ is strictly weaker than the constraint $Y \geqslant 0$.

The first relaxation we will focus on is a **manifold optimization** relaxation of k-means in Chapter 3. First note that k-means can be seen as an optimization problem in a discrete set (the constraint set of (1.3)), but if we remove the non-negative constraint $Y \geqslant 0$ we obtain a compact manifold. We consider the relaxation of (1.3) where we relax the non-negative constraint $Y \geqslant 0$ to a penalization in the objective, and restrict the minimization to $Y \in M$ where M is a smooth submanifold of $\mathbb{R}^{n\times k}$:

$$\begin{aligned} &\text{minimize} \quad \operatorname{Tr}(DYY^\top) + \lambda\|Y_-\|_F^2 \\ &\text{subject to} \quad Y \in M. \end{aligned} \tag{1.6}$$

Here Y_- indicates the negative entries of Y, λ is a non-negative parameter that penalizes Y with negative entries, and M is the submanifold

$$M = \{Y \in \mathbb{R}^{n\times k} : \; Y^\top Y = I_k, \; YY^\top 1 = 1\}. \tag{1.7}$$

By removing $Y \geqslant 0$ from the constraint set, our discrete feasible set becomes a smooth manifold without boundary, so we can use manifold optimization algorithms to solve the problem.

Also note that adding the constraint $YY^\top 1 = 1$ to spectral clustering is simple and doesn't change its spectral nature (in particular, if $\lambda = 0$ the solution can be computed from the top $k-1$ eigenvectors of the projection of D onto $\{1\}^\perp \subset \mathbb{R}^n$). What makes this optimization significantly different from spectral clustering is the term $\lambda\|Y_-\|_F^2$ in the objective.

In Chapter 3 we explain how to implement an efficient manifold optimization algorithm to approach problem (1.6) and we provide numerical

experiments that suggest that, in some settings, the algorithm converges to the optimal solution of (k-means). Unfortunately we do not have theoretical guarantees for this algorithm yet (the objective of (1.6) is not convex and the algorithm may converge to local minima). However, we will be able to combine this efficient algorithm with the proofs from Chapter 4 to provide an efficient algorithm with a certificate of optimality in Chapter 5.

1.4.2 Exact recovery of clustering solutions using convex relaxations

We consider three different convex relaxations of the k-medians and k-means objectives, described in Figures 1.2, 1.3, and 1.4.

(i) A semidefinite programming (SDP) relaxation of k-means introduced by Peng and Wei [56],

(ii) a linear programming (LP) relaxation of k-means,

(iii) and a standard linear programming (LP) relaxation of k-medians,

See Section 2.1.1 for a brief background in linear and semidefinite programs.

We provide deterministic conditions that if satisfied by the point set P, imply that the corresponding convex optimization program is tight (and therefore it recovers the exact solution of problems (k-medians) or (k-means)).

The deterministic conditions we find do not provide geometric intuition a priori. Therefore, in order to evaluate their expressivity, we consider a random point model of naturally clustered data introduce by Nellore and Ward [50] known as the *stochastic ball model* depicted in Figure 1.5. The

$$\begin{aligned} \underset{X \in \mathbb{R}^{N \times N}}{\text{minimize}} \quad & \frac{1}{2}\operatorname{trace}(DX) && \text{(k-means sdp)} \\ \text{subject to} \quad & X1 = 1,\ \operatorname{trace}(X) = k,\ X \geqslant 0,\ X \succeq 0. \end{aligned}$$

Figure 1.2: Peng and Wei's semidefinite programming relaxation of (k-means)
The symmetric matrix D is defined as $D_{ij} := \|x_i - x_j\|^2$ for $x_i, x_j \in P$, and $X \succeq 0$ means that X is symmetric and positive semidefinite. If $A_1, \ldots A_k$ is a cluster, the corresponding projection matrix X is $\sum_{t=1}^{k} \frac{1}{|A_t|} 1_{A_t} 1_{A_t}^\top$ where the indicator vector $(1_{A_t})_i$ is 1 if $x_i \in A_t$ and 0 otherwise. Note that relaxation (k-means sdp) relaxes the set of all cluster projection matrices into a subset of the positive semidefinite matrices.

$$\begin{aligned} \underset{X \in \mathbb{R}^{N \times N}}{\text{minimize}} \quad & \frac{1}{2}\operatorname{trace}(DX) && \text{(k-means lp)} \\ \text{subject to} \quad & X1 = 1,\ \operatorname{trace}(X) = k,\ X = X^\top,\ X_{ii} \geqslant X_{ij}\ \forall i, j \in [n],\ X_{ij} \geqslant 0. \end{aligned}$$

Figure 1.3: Linear programming relaxation of (k-means)
This linear programming relaxation replaces the semidefinite constraint from (k-means sdp) with looser linear constraints. In general, linear programs are numerically more efficient and simpler to analyze than semidefinite programs. However we prove the quality of the solution of (k-means lp) is inferior to the one of (k-means sdp).

$$\begin{aligned} \underset{z \in \mathbb{R}^{N \times N},\, y \in \mathbb{R}^N}{\text{minimize}} \quad & \sum_{i=1}^{n} \sum_{j=1}^{n} d(x_i, x_j) z_{ij} && \text{(k-medians lp)} \\ \text{subject to} \quad & \sum_{i=1}^{n} z_{ij} = 1\ \forall j \in [n],\ \ z_{ij} \leqslant y_i\ \forall i, j \in [n], \\ & \sum_{i=1}^{n} y_i = k,\ \ z_{ij}, y_i \in [0, 1]. \end{aligned}$$

Figure 1.4: Linear programming relaxation of (k-medians)
The relaxation (k-medians lp) consists of replacing the discrete set $\{0, 1\}$ by the interval $[0, 1]$. In the original optimization problem (k-medians) the variable y_i indicates whether the point x_i is a center or not, while z_{ij} is 1 if the point x_j is assigned to x_i as center, and 0 otherwise. The solution for the integer programming problem (where $\{z_{ij}, y_i\} \in \{0, 1\}$) corresponds to the adjacency matrix for a graph consisting of disjoint star-shaped graphs like the one shown in Figure 1.1.

premise behind evaluating algorithms in this model is that a good algorithm should recover the right clusters when the solution is *obvious*.

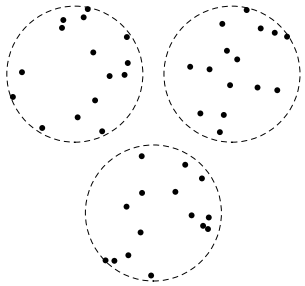


Figure 1.5: Stochastic ball model.
Example of an instance of the stochastic ball model in $\mathbb{R}^2$. Here $\mathcal{D}$ is the uniform distribution in the unit ball, $k = 3$, $n = 15$, and $\Delta = 2.2$.

Definition 1.4.1 ($(\mathcal{D}, \gamma, n)$-stochastic ball model)**.** Let $\{\gamma_a\}_{a=1}^k$ be ball centers in $\mathbb{R}^m$. For each a, draw i.i.d. vectors $\{r_{a,i}\}_{i=1}^n$ from some rotation-invariant distribution $\mathcal{D}$ whose support is the unit ball. The points from cluster a are then taken to be $x_{a,i} := r_{a,i} + \gamma_a$. We denote $\Delta := \min_{a \neq b} \|\gamma_a - \gamma_b\|_2$.

Note that when $\Delta < 2$ the the clusters overlap and the "cluster solution" is no longer well-defined. We now present informal statements of our main results; see specific sections for more details.

Theorem 1.4.1. *Under $(\mathcal{D}, \gamma, n)$-stochastic ball model and with high probability, Peng and Wei's SDP relaxation of* k*-means* (k-means sdp) *recovers the clusters up to separation $\Delta > \min\{2\sqrt{2(1 + 1/m)},\ 2 + k^2/m\}$.*

Theorem 1.4.2. *Under the $(\mathcal{D}, \gamma, n)$-stochastic ball model a simple LP relaxation for the* k*-means objective* (k-means lp) *with high probability fails to recover the*

exact clusters at separation $\Delta < 4$, even for $k = 2$ clusters.

Theorem 1.4.3. *For any constant $\epsilon > 0$, there exists n sufficiently large so that the k-medians LP relaxation* (k-medians lp) *is tight and recovers the true clustering of the points under the $(\mathcal{D}, \gamma, n)$-stochastic ball model with arbitrarily high probability as long as $\Delta > 2 + \epsilon$.*

The proofs of Theorems 1.4.3, 1.4.2, and 1.4.1 use the same general technique. First, using convex duality, we provide deterministic conditions on the data under which the convex optimization program is *tight* (meaning, the solution of the respective relaxation coincides with the globally optimal partition). We find those deterministic conditions using a technique known as *dual certificate* described in Section 2.1.1.1. Using random matrix theory we prove that under the stochastic ball model, the deterministic conditions hold with high probability provided that the separation between the centers is not too small.

Table 1.1 summarizes the state of the art for recovery guarantees under the stochastic ball model. Theorem 1.4.3 is an improvement over [50], where it was shown that (k-medians lp), with high probability, recovers clusters drawn from the stochastic ball model provided the smallest distance between ball centers is $\Delta \geqslant 3.75$. We know that exact recovery only makes sense for $\Delta > 2$ (i.e., when the balls are disjoint). Once $\Delta > 4$, any two points within a particular cluster are closer to each other than any two points from different clusters, and so in this regime, cluster recovery follows from a simple distance thresholding.

Theorems 1.4.3 and 1.4.2 are tight in their dependence on the cluster separation Δ and appear in [8]. Theorem 1.4.1 is proven through two different dual certificates, the first bound corresponds to the dual certificate from [8] and the second bound comes from the certificate from [34]. Neither of this bounds is tight and in some regimes the certificate from [8] gives a better guarantee than the certificate from [34] whereas in other regimes the opposite is true. The question of what is an optimal dual certificate remains open for this problem. An answer to this question could arise from comparing both certificates with the pre-certificate defined in [62].

Under the assumptions of the theorems above, popular heuristic algorithms such as *Partitioning around Medoids* (PAM) and *Lloyd's algorithm* (for k-medians and k-means, respectively) can fail with high probability. Even with arbitrarily large cluster separation, variants of Lloyd's algorithm, such as k-means++ with overseeding by any constant factor, fail with high probability at exact cluster recovery. See Figure 1.6 for an illustration and [8] for details.

In our numerical experiments we observed that the k-medians linear program (k-medians lp) is often tight, even when the data points are drawn from a single spherical gaussian, were no cluster structure is expected. It remains an open problem to understand this phenomenon [10]. The k-means semidefinite relaxation (k-means sdp) however, is not tight for more general data models, like mixtures of subgaussian distributions. In Section 1.4.4 we describe an algorithm that involves solving the SDP and rounding the ob-

Method	**Sufficient Condition**	**Optimal?**	**Reference**
Thresholding	$\Delta > 4$	Yes	(simple exercise)
k-medians LP	$\Delta \geqslant 4$	No	Theorem 2 in [27]
	$\Delta \geqslant 3.75$	No	Theorem 1 in [50]
	$\Delta > 2$	Yes	Theorem 4.2.3
k-means LP	$\Delta > 4$	Yes	Theorem 4.3.2
k-means SDP	$\Delta > 2\sqrt{2(1+1/m)}$	No	Theorem 4.1.4
	$\Delta > 2 + k^2/m$	No	Theorem 4.1.9
Spectral k-means ($k = 2$)	$\Delta > \Delta^\star$	Yes	Theorem 14 in [34]

Table 1.1: Summary of cluster recovery guarantees under the stochastic ball model.

The second column reports sufficient separation between ball centers in order for the corresponding method to provably give exact recovery with high probability. The third column reports whether the sufficient condition on Δ cannot be improved. Here, $\Delta^\star = \Delta^\star(\mathcal{D}, k)$ denotes the smallest value for which $\Delta > \Delta^\star$ implies that minimizing the k-means objective recovers planted clusters under the $(\mathcal{D}, \gamma, n)$-stochastic ball model with probability $1 - e^{-\Omega_{\mathcal{D},\gamma}(n)}$. In [34] we prove the surprising result that $\Delta^\star > 2$ at least in dimension $m \leqslant 2$.

Figure 1.6: Failure of Lloyd's algorithm.
Recall the steps 1-3 in Lloyd's algorithm. The output of the algorithm depends on its initialization. For example, let us say we want to cluster points drawn from the stochastic ball model, illustrated in this figure. If the initial guess has only one point from the two balls at the left and two points from the ball in the right, then Lloyd's algorithm will fail to identify the correct clusters, obtaining an output similar to the one depicted in this figure. The probability of having a bad initial guess is positive and grows exponentially in k.

tained solution to a partition and we provide approximation guarantees for the algorithm.

1.4.3 Fast certification of k-means optimality

On one hand we have very fast clustering algorithms like Lloyd's [43] or manifold optimization based algorithms, whose solutions may be far from optimal. On the other hand we have optimization based algorithms like k-means SDP, which are slow but provide a certificate of optimality. What if we could combine the best of both worlds and obtain a fast algorithm with a certificate of optimality?

Recently Bandeira devised a general technique to efficiently provide

certifiably correct solutions to hard problems [9]. This technique leverages three components:

(i) A fast non-convex solver that produces the optimal solution with high probability (under some probability distribution of problem instances).

(ii) A convex relaxation that is tight with high probability (under the same distribution).

(iii) A fast method of computing a certificate of global optimality for the output of the non-convex solver in (i) by exploiting convex duality with the relaxation in (ii).

Using Bandeira's technique in Chapter 5 we develop a quasi-linear time algorithm that provides certificates of k-means optimality of clusters [34], where (i) and (ii) are chosen to be k-means++ and k-means SDP respectively.

In many useful applications the k-means SDP is not tight. In fact, in order for Bandeira's technique to have practical value, we need to develop a *robust* version of it. In particular, a version that works even when the relaxation is not tight. This is an open problem that basically requires an algorithm that given an approximation solution of a convex optimization problem, it provably provides an approximate solution of the dual problem (faster than solving the dual problem).

The importance of this problem goes beyond clustering applications. It could provide a practical way of measuring the quality of solutions found by fast but maybe unreliable methods. For large datasets, problems tend to

be intractable for higher precision methods and this certificate can be of practical relevance.

1.4.4 Approximation guarantees

We earlier discussed that (k-means sdp), the semidefinite relaxation of k-means, recovers the optimal clusters for the stochastic ball model. In Chapter 6 we study its performance under the general subgaussian mixture model, which includes the stochastic ball model and the Gaussian mixture model as special cases.

The semidefinite program is not typically tight under this general model, but the optimizer can be interpreted as a denoised version of the data and can be rounded in order to produce a good estimate for the centers (and therefore produce a good clustering).

To see this, let P denote the $m \times N$ matrix whose columns are the coordinates of the points we want to cluster: $\{x_{t,i}\}_{t\in[k],i\in[|A_t|]}$. Notice that whenever the semidefinite relaxation is tight, X has the form (1.8),

$$X_{ij} = \begin{cases} \frac{1}{|A_t|} & \text{if } i,j \in A_t \\ 0 & \text{otherwise} \end{cases} \tag{1.8}$$

then for each $t \in [k]$, PX has $|A_t|$ columns equal to the centroid of points assigned to A_t.

In particular, if X is k-means-optimal, then PX reports the k-means-optimal centroids (with appropriate multiplicities). Next, we note that every SDP-feasible matrix $X \geqslant 0$ satisfies $X^\top 1 = X1 = 1$, and so $X^\top$ is a stochas-

tic matrix, meaning each column of PX is still a weighted average of columns from P. Intuitively, if the SPD relaxation (k-means sdp) were close to being tight, then the SDP-optimal X would make the columns of PX close to the k-means-optimal centroids. Empirically, this appears to be the case (see Figure 1.7 for an illustration). Overall, we may interpret PX as a denoised version of the original data P, and we leverage this strengthened signal to identify good estimates for the k-means-optimal centroids.

What follows is a summary of our relax-and-round procedure for (approximately) solving the k-means problem (k-means):

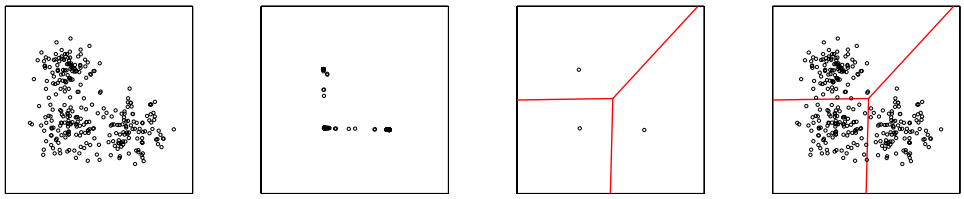
Relax-and-round k-means clustering procedure.
Given and $m \times N$ data matrix $P = [x_1 \cdots x_N]$, do:

(i) Compute distance-squared matrix D defined by $D_{ij} = \|x_i - x_j\|_2^2$.

(ii) Solve (k-means sdp), resulting in optimizer X.

(iii) Cluster the columns of the denoised data matrix PX.

For step (iii), we find there tends to be k vectors that appear as columns in PX with particularly high frequency, and so we are inclined to use these as estimators for the k-mean-optimal centroids (see Figure 1.7, for example). Running Lloyd's algorithm for step (iii) also works well in practice. To obtain theoretical guarantees, we instead find the k columns of PX for which the unit balls of a certain radius centered at these points in $\mathbb{R}^m$ contain the



(a) (b) (c) (d)

Figure 1.7: Illustration of the relax and round k-means clustering procedure **(a)** Draw 100 points at random from each of three spherical Gaussians over $\mathbb{R}^2$. These points form the columns of a 2×300 matrix P. **(b)** Compute the 300×300 distance-squared matrix D from the data in (a), and solve the k-means semidefinite relaxation (k-means sdp) using SDPNAL+v0.3 [71]. (The computation takes about 16 seconds on a standard MacBook Air laptop.) Given the optimizer X, compute PX and plot the columns. We interpret this as a denoised version of the original data P. **(c)** The points in (b) land in three particular locations with particularly high frequency. Take these locations to be estimators of the centers of the original Gaussians. **(d)** Use the estimates for the centers in (c) to partition the original data into three subsets, thereby estimating the k-means-optimal partition.

most columns of PX (see Theorem 6.5.1 for more details). An implementation of our procedure is available on GitHub [67] and an interactive web visualization of the MNIST numerical simulation is available on [66].

In Chapter 6 we provide performance guarantees for the k-means semidefinite relaxation (k-means sdp) when the point cloud is drawn from a subgaussian mixture model. We adapt ideas from Guédon and Vershynin [32] and obtain approximation guarantees comparable with the state of the art for learning mixtures of Gaussians despite the fact that our algorithm is a generic k-means solver and uses no model assumptions. In Section 6.5.1 we illustrate its numerical performance on the MNIST handwritten data set.

Recent work by Yan and Sarkar [70] adapted a similar version of our algorithm for kernel matrices and proved its strong consistency. They also prove that spectral methods are only weakly consistent, which provides some theoretical evidence of what we observe numerically: the semidefinite programming relaxation of k-means performs much better than other algorithms at finding the k-means solution.

We summarize our approximation result in the following theorem

Theorem 1.4.4. *Given $x_1, \ldots, x_N$ points drawn independently from a mixture of k subgaussian distributions in $\mathbb{R}^m$. Say that the subgaussian a, for $1 \leqslant a \leqslant k$ has center γ_a, and σ^2 is an upper bound on the maximum covariance. Let $\Delta_{\min} = \min_{a \neq b} \|\gamma_a - \gamma_b\|$ and similarly $\Delta_{\max}$. If $k\sigma \lesssim \Delta_{\min} \leqslant \Delta_{\max} \lesssim K\sigma$, then we have that there exists a permutation π on $\{1, \ldots, k\}$ such that*

$$\frac{1}{k} \sum_{i=1}^{k} \|\nu_i - \tilde{\gamma}_{\pi(i)}\|_2^2 \lesssim k K^2 \sigma^2, \tag{1.9}$$

where ν_i is what our algorithm chooses as the ith center estimate and $\tilde{\gamma}_a$ is the average of the points sampled from the subgaussian a.

1.4.5 A polynomial-time relaxation for the Gromov-Hausdorff distance

In the previous sections we have summarized how to use convex relaxations (i) to find exact solutions to clustering problems, (ii) to certify optimality of solutions acquired by faster but sometimes unreliable algorithms, and (iii) to provide approximate solutions with explicit approxima-

tion bounds via a relax and round procedure. Now we consider a different approach to optimization problems where we relax but do not round.

The Gromov-Hausdorff distance between finite metric spaces X and Y, introduced in Section 1.2, can be formulated as an NP-hard optimization problem that for now we write in the abstract form (1.10). In Chapter 7 we consider a semidefinite relaxation of (1.10) obtaining a tractable optimization problem of the form (1.11).

$$\begin{aligned} d_{GH}(X,Y) := \text{minimize} \quad & f_{X,Y}(z) \\ \text{subject to} \quad & z \in \mathcal{S}. \end{aligned} \tag{1.10}$$

$$\begin{aligned} \tilde{d}_{GH}(X,Y) := \text{minimize} \quad & \tilde{f}_{X,Y}(z) \\ \text{subject to} \quad & z \in \tilde{\mathcal{S}}. \end{aligned} \tag{1.11}$$

The relaxation (1.11) defines $\tilde{d}_{GH}$, which we prove is a pseudometric on point clouds and can be computed in polynomial time. We also show $\tilde{d}_{GH}$ is a lower bound for the Gromov-Hausdorff distance d_{GH}. Our semidefinite relaxation (1.11) also provides $z \in \tilde{\mathcal{S}}$ that can be interpreted as a relaxed correspondence between point clouds.

We study the topological properties of the relaxed pseudodistance $\tilde{d}_{GH}$ (like convergence and compactness) and we observe that for almost every space X there exists a small local neighborhood where the metrics d_{GH} and $\tilde{d}_{GH}$ are equivalent (see Corollary 7.2.8 and previous definitions).

In Section 7.3 we exploit the theoretical observations to propose a non-convex optimization algorithm to approach the registration problem efficiently. The output of this algorithm not only provides a local optimum for the registration problem, but also an upper bound for the Gromov-Hausdorff distance.

The work in Chapter 7 appears in [68]. Note that a similar version to our SDP was recently introduced in [38] and further studied in [45] and [24]. Our work provides theoretical validation for some of the computational phenomena observed therein and complements their theoretical framework.

Chapter 2

Background

"The great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

– R. Tyrrell Rockafellar, in SIAM Review, 1993.

2.1 Optimization

In this section we present a small summary of optimization concepts we use in this thesis. For more comprehensive background information we refer the reader to the classic texts in convex optimization [51, 16] and [4] for manifold optimization.

Consider a general optimization problem of the form

$$\begin{aligned} &\text{minimize} && f_D(x) \\ &\text{subject to} && x \in \mathcal{S}. \end{aligned} \tag{2.1}$$

When dealing with finite sets of data, we generally can formulate the problems as a combinatorial optimization problem, where $\mathcal{S}$ is a discrete set, and our problem consists of finding an optimal object among a finite set.

Some combinatorial optimization problems can be solved efficiently

with specialized algorithms, but a large family of them, including the ones studied in this thesis, are NP-hard (and therefore one cannot expect to find an efficient algorithm that solves the problem in general). A useful strategy in this case is the relax-and-round paradigm introduced in Section 1.3.

In the convex optimization setting one relaxes the problem (2.1) to a minimization of a convex function over a convex set. In particular we focus in in linear programming (LP) and semidefinite programming (SDP). For LP the convex set considered is a convex polytope (i.e. the intersection of half spaces in Euclidean space). For SDP, the convex set is a spectahedron (i.e. the intersection of the cone of positive semidefinite matrices with an affine space). Both LP and SDP are particular cases of conic optimization, which we describe in Section 2.1.1. Conic optimization problems can be efficiently solved with interior point methods (and in general algorithms for LP tend to be more efficient than algorithms for SDP [51]). Conic optimization problems have an advantage with respect to generic convex optimization problems: conic problems have a dual problem that can be easily expressed in closed form. The dual problem is very useful to provide algorithms and theoretical results

A recently popular alternative to convex optimization is manifold optimization [4], where the set $\mathcal{S}$ is relaxed to a smooth convex manifold and the geometry of the manifold is exploited to obtain efficient algorithms. The main advantage of manifold optimization algorithms with respect to convex optimization is that for reasonably nice manifolds, manifold opti-

mization algorithms tend to run and converge much faster than interior point methods. The disadvantage is that in general they converge to local optima.

2.1.1 Cone programming

A set $K \subset \mathbb{R}^n$ is a cone if $x \in K$ implies $tx \in K$ for all $t \geqslant 0$. Let $K \subset \mathbb{R}^n$ and $L \subset \mathbb{R}^m$ be closed convex cones, consider $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and let $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear operator. Then a cone programming problem is an optimization problem of the form (P).

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & -\langle c, x\rangle \\ \text{subject to} \quad & b - Ax \in L \\ & x \in K \end{aligned} \tag{P}$$

For K closed convex cone, we define its dual cone as K^* as

$$K^* := \{y : \langle y, x\rangle \geqslant 0 \quad \forall x \in K\}.$$

Then the dual problem of (P) is defined as (D):

$$\begin{aligned} \underset{y}{\text{maximize}} \quad & -\langle b, y\rangle \\ \text{subject to} \quad & A^*y - c \in K^* \\ & y \in L^* \end{aligned} \tag{D}$$

where A^* denotes the adjoint of A, while K^* and L^* denote the dual cones of K and L, respectively.

In the optimization jargon we say that (P) is the primal problem and (D) is the dual problem. We say (P) has objective function $x \mapsto -\langle c, x\rangle$, and a point $x \in \mathbb{R}^n$ is said to be feasible for (P) if it satisfies the constraints in (P) (i.e. $b - Ax \in L$ and $x \in K$).

Proposition 2.1.1 (Weak duality)**.** *Let* x *and* y *be feasible points for* (P) *and* (D) *respectively. Then* $-\langle b, y\rangle \leqslant -\langle c, x\rangle$*.*

Proof. Since $b - Ax \in L$ and $y \in L^*$ then the definition of dual cone implies

$$0 \leqslant \langle b - Ax, y\rangle = \langle b, y\rangle - \langle A^* y, x\rangle$$

then $-\langle b, y\rangle \leqslant -\langle A^* y, x\rangle$. The same computation with $x \in K$ and $A^* y - c \in K^*$ gives $-\langle A^* y, x\rangle \leqslant -\langle c, x\rangle$ which gives the result. □

Weak duality says that the dual problem provides lower bounds for the primal objective. Strong duality says that the optimal value of (P) actually equals the optimal value of (D) (see [51] for a proof).

Theorem 2.1.2 (Strong duality)**.** *The problem* (P) *is feasible and has bounded optimal value* α *if and only if* (D) *is feasible and has bounded optimal value* α*.*

2.1.1.1 Complementary slackness and dual certificates

If x and y are feasible for (P) and (D) respectively, weak duality implies

$$-\langle y, b\rangle \leqslant -\langle y, Ax\rangle = -\langle A^* y, x\rangle \leqslant -\langle c, x\rangle$$

or equivalently, $\langle c - A^* y, x\rangle \leqslant 0 \leqslant \langle y, b - Ax\rangle$. By strong duality, x and y are optimal if and only if these inequalities are equal. That is,

$$\langle A^\top y - c, x\rangle = 0 = \langle y, b - Ax\rangle.$$

In that sense, the primal variable x is complementary to the dual constraint $A^\top y - c$ just as the dual variable y is complementary to the primal constraint $b - Ax$. These orthogonality relations are sometimes helpful when expressing the optimal y (called the dual certificate) in terms of the optimal x.

An interesting interpretation for the term *dual certificate* is that given x feasible for (P), if one can find y feasible for (D) such that $-\langle c, x\rangle = -\langle b, y\rangle$ (or equivalently $\langle A^\top y - c, x\rangle = 0 = \langle y, b - Ax\rangle$), then y is a proof of x's optimality for (P).

2.1.2 Manifold optimization

Let us consider an optimization problem of the form

$$\begin{aligned} &\text{minimize} && f(Y) \\ &\text{subject to} && Y \in M, \end{aligned} \tag{2.2}$$

where $f : M \to \mathbb{R}$ is a smooth function and M is a compact Riemannian manifold.

For this kind of problems there is a beautiful theory [4] that allows us to think of the optimization problem (2.2) as an unconstrained optimization

where we replace the usual Euclidean ambient space by the Riemannian manifold M.

The basic gradient descent algorithm relies on gradient and retraction functions,

$$\operatorname{grad}_f : M \to TM, \tag{2.3}$$

$$\operatorname{retr}_Y : T_Y M \to M. \tag{2.4}$$

The gradient is computable using the Riemannian structure. The retraction is a choice of map which should satisfy

$$\operatorname{retr}_Y(0) = 0, \quad \left.\frac{d}{dt}\right|_{t=0} \operatorname{retr}_Y(tV) = V. \tag{2.5}$$

A canonical choice of retraction map is the exponential map for M, but this is not always computationally feasible. If M is a submanifold of euclidean space, $Y \in M$ and $V \in T_Y M$, then $\operatorname{retr}_Y(V)$ will be a first order approximation to $Y + V$.

The algorithm consists of iteratively following the gradient of f in the tangent space and then retracting back into the manifold:

$$Y_{n+1} = \operatorname{retr}_{Y_n}(-\alpha_n \operatorname{grad}_f(Y_n)).$$

The stepsize α_n can be set to be a small constant or adaptively chosen through a line search. Second order algorithms like trust regions have also been adapted to the manifold optimization setting [4].

In this thesis we restrict ourselves to first order methods, where gradient descent methods with backtracking Amijo line-search are proven to converge to a stationary point under mild hypotheses [13].

Theorem 6 in [13]. *Let* M *a Riemannian manifold and* $f : M \to \mathbb{R}$ *bounded from below. Assume that* $f \circ \mathrm{retr}_Y$ *is Lipchitz with constant* L *independent of* Y*. Then a gradient descent on* M *with backtracking Amijo line-search initialized at* Y_0 *returns* Y_* *such that*

$$f(Y_*) \leqslant f(Y_0) \quad \textit{and} \quad \|\mathrm{grad}_f(Y_*)\| \leqslant \epsilon$$

in $O(1/\epsilon^2)$ *iterations.*

Chapter 3

Manifold optimization techniques for k-means clustering

3.1 The k-means manifold.

Recall our manifold optimization relaxation of k-means (3.1) introduced in Section 1.4.1

$$\begin{aligned} &\text{minimize} \quad \operatorname{Tr}(DYY^\top) + \lambda \|Y_-\|_F^2 \\ &\text{subject to} \quad Y \in M, \end{aligned} \tag{3.1}$$

where λ is a non-negative parameter, Y_- indicates the negative entries of Y, and M is the submanifold

$$M = \{Y \in \mathbb{R}^{n\times k} : \, Y^\top Y = I_k, \, YY^\top 1 = 1\}. \tag{3.2}$$

The relaxation (3.1) is a constrained optimization where the set of constraints is a Riemannian manifold, so we can use the theory described in Section 2.1.2.

This chapter is based on the publication:
Timothy Carson, Dustin G. Mixon, Soledad Villar, Rachel Ward. *Manifold optimization for k-means clustering* Proceedings of the 2017 International Conference on Sampling Theory and Applications (SampTA), 2017 (to appear)
The author contributed by designing the algorithm and its implementation.

In order to implement the manifold optimization relaxation of k-means we need to explicitly construct the gradient and retraction maps (2.3) and (2.4). The tangent space to M at Y is given by

$$T_Y M = \{V \in \mathbb{R}^{n\times k} : V^\top Y + Y^\top V = 0, (VY^\top + YV^\top)1 = 0\}. \quad (3.3)$$

Our manifold is a submanifold of a Euclidean space, and our objective function is defined on the entirety of this Euclidean space. As such, we may compute the gradient of the objective function on our manifold by orthogonally projecting its gradient in Euclidean space onto the tangent space to our manifold. That is, from the orthogonal projection $\Pi_{T_Y M} : T_Y\mathbb{R}^{n\times k} \to T_Y M$ we can compute

$$\operatorname{grad}_f^M(Y) = \Pi_{T_Y M} \circ \nabla f(Y)$$

where ∇f is the gradient of f in the ambient Euclidean space $\mathbb{R}^{n\times k}$. For our objective function (with parameter λ),

$$f_\lambda(Y) = \operatorname{Tr}(DYY^\top) + \lambda\|Y_-\|^2,$$

the gradient is computed in Section 3.2 to be

$$\nabla f_\lambda(Y) = 2DY + 2\lambda(Y)_-. \quad (3.4)$$

In Section 3.3 we compute the orthogonal projection. It is:

$$\Pi_{T_Y M}(W) = W - 2Y\Omega - (x1^\top + 1x^\top)Y, \quad (3.5)$$

where

$$x = \frac{1}{n} W Y^\top 1 \in \mathbb{R}^n,$$
$$\Omega = \frac{1}{4}(W^\top Y + Y W^\top - 2Y^\top (x1^\top + 1x^\top)Y) \in \mathbb{R}^{k\times k}.$$

We use the following retraction:

$$\mathrm{retr}_Y(V) = \exp(B)\exp(A')Y, \tag{3.6}$$

where

$$A = Y^\top V \in \mathbb{R}^{k\times k},$$
$$A' = YAY^\top \in \mathbb{R}^{n\times n},$$
$$B = VY^\top - YV^\top - 2A' \in \mathbb{R}^{n\times n}.$$

Here exp denotes the matrix exponential. We explain this retraction in Section 3.6.

3.2 Gradient of the objective function

We compute the ambient space gradient $\nabla f_\lambda(Y)$ of f_λ. By definition we know $\nabla f_\lambda(Y) = W$ if and only if for all $V \in T_Y M$ we have

$$\langle V, W\rangle = D_Y f_\lambda(V) = \mathrm{Tr}(D(VY^\top + YV^\top)) + \lambda\,\mathrm{Tr}(V(Y^\top)_- + (Y_-)V^\top)$$

where $D_Y f_\lambda(V)$ is the directional derivative of f_λ. Equivalently,

$$\mathrm{Tr}(WV^\top) = \mathrm{Tr}(((D + D^\top)Y + 2\lambda(Y_-))V^\top).$$

Since D is symmetric we find (3.4).

3.3 Projection of a vector onto $T_Y\mathcal{M}$

Let $L_1 : \mathbb{R}^{n\times k} \to \mathbb{R}^{k\times k}_{sym}$ be $L_1(W) = W^\top Y + Y^\top W$ and let $L_2 : \mathbb{R}^{n\times k} \to \mathbb{R}^n$ be $L_2(W) = (WY^\top + YW^\top)1$. We can write the tangent space as $T_Y\mathcal{M} = \ker(L)$ where $L = L_1 \oplus L_2 : \mathbb{R}^{n\times k} \to \mathbb{R}^{k\times k}_{sym} \times \mathbb{R}^n$.

We can use $\ker(L)^\perp = \mathrm{im}(L^*)$ to compute a parameterization for $(T_Y\mathcal{M})^\perp$. Then we will solve $W - L^*(\Omega, x) \in \ker(L)$ for (Ω, x) to find the projection $\Pi_{T_Y\mathcal{M}}(W) = W - L^*(\Omega, x)$.

We calculate that for Ω symmetric:

$$\langle L_1 W, \Omega\rangle = \langle W^\top Y, \Omega\rangle + \langle Y^\top W, \Omega\rangle = \langle W^\top, \Omega Y^\top\rangle + \langle W, Y\Omega\rangle = 2\langle W, Y\Omega\rangle,$$

from which we see $L_1^*\Omega = 2Y\Omega$. Now calculate for $x \in \mathbb{R}^n$:

$$\langle L_2 W, x\rangle = \langle (WY^\top + YW^\top)1, x\rangle = \langle WY^\top + YW^\top, x1^\top\rangle = \langle W, x1^\top Y + 1x^\top Y\rangle,$$

so $L_2^* x = (x1^\top + 1x^\top)Y$

Now we can find Ω and x so that $W - L_1^*\Omega - L_2^* x \in \ker(L)$ by solving the system of equations:

$$\begin{cases} L_1(W - L_1^*\Omega - L_2^* x) = 0 \\ L_2(W - L_1^*\Omega - L_2^* x) = 0 \end{cases}$$

The first equation reads

$$W^\top Y + Y^\top W - 4\Omega - 2Y^\top(x1^\top + 1x^\top)Y = 0$$

We can use this to substitute Ω in the second equation to get:

$$u + Bx = 0$$

where $u = (I_n - YY^\top)WY^\top 1$ and $B = -n(I_n - YY^\top)$. In particular we can choose x and Ω as below (3.5). (There is nonuniqueness in x and Ω because the image of L is not the full stated range, but of course the projection is unique.)

3.4 Homogenous structure of M

Recall the Definition (3.2) of M. Let M_0 be the manifold

$$M_0 = \{YY^\top : Y \in M\} \subset \mathbb{R}^{n\times n}_{sym}.$$

The manifold M_0 is the set of orthogonal projections onto a k dimensional subspace of $\mathbb{R}^n$ including the vector 1_n, and as such each member of M_0 is determined by its image. A point in the manifold M has the additional information of a choice of basis of the image of $X = YY^\top$.

For a subspace $A \subset \mathbb{R}^n$, $O(A)$ is the group of orthogonal matrices for which $Av = v$ for all $v \in A^\perp$. Let $\mathbb{P} = \{1_n\}^\perp \subset \mathbb{R}^n$. We can see M as a homogenous space; it has a transitive action by $O(\mathbb{P}) \times O(\mathbb{R}^k)$ given by multiplication by the first factor on the left and the $O(\mathbb{R}^k)$ factor on the right:

$$\begin{aligned} M \times O(\mathbb{P}) \times O(\mathbb{R}^k) &\to M \\ (Y, Q, R) &\mapsto QYR. \end{aligned}$$

The multiplication on the right by an element of $O(\mathbb{R}^k)$ controls changes which change Y but not X, which may be seen directly from the computation

$(YR)(YR)^\top = YRR^\top Y^\top = YY^\top$. The multiplication on the left by $Q \in O(\mathbb{P})$ allows for any change in $X \in M_0$.

Multiplication of $Y \in M$ on the right by $R \in O(\mathbb{R}^k)$ is always equivalent to multiplication of Y on the left by $R' = YRY^\top$:

$$R'Y = (YRY^\top)Y = YR(Y^\top Y) = YR.$$

The matrix $(YRY^\top)$ is an orthogonal projection onto $\mathrm{im}(X)$ composed with an orthogonal transformation of $\mathrm{im}(X)$, which may also be shown by computing

$$R'(I - X) = 0, \quad R(R')^\top = X.$$

Recalling that X is an orthogonal projection, the first equality shows that R' annihilates $\mathrm{im}(X)^\perp$ and the second equality shows that R' acts as an orthogonal transformation of $\mathrm{im}(X)$ (on which X is the identity).

For each Y_0 the action by $O(\mathbb{P}) \times O(\mathbb{R}^k)$ has a stabilizer which is determined by $X_0 = Y_0 Y_0^\top$. This is due to redundancies of the right multiplication in the left multiplication. The action by $O(\mathbb{R}^k)$ generates all $Y \in M$ with the same X_0:

$$\{Y_0 R : R \in O(\mathbb{R}^k)\} = \{Y \in M : YY^\top = Y_0 Y_0^\top\},$$

but there are also elements of $O(\mathbb{P})$ which fix X_0 namely,

$$\begin{aligned}&\{Q \in O(\mathbb{R}^n) : QX_0 = X_0 Q, Q1_n = 1_n\} \\ &\quad = O(\mathrm{im}(X)) \oplus O(\ker(X)) \subset O(\mathbb{P}) = \{Q \in O(\mathbb{R}^n) : Q1_n = 1_n\}.\end{aligned}$$

3.5 Splitting the tangent space to M

We may use our understanding of M as a homogenous space to compute a splitting of the tangent space $T_Y M$ into two orthogonal parts; those which generate changes which fix X, and its perpendicular space. Let $\mathfrak{so}(\mathbb{R}^n)$ be the set of antisymmetric matrices in $\mathbb{R}^n$.

The matrices in $\{CY : C \in \mathfrak{so}(\mathbb{R}^n)\}$ which are tangent to the direction of fixed X (generated by $O(\mathrm{im}(X)) \oplus O(\ker(X))$) are

$$\{C \in \mathfrak{so}(\mathbb{R}^n) : CX = XC\}.$$

This is the kernel of the linear map $\mathfrak{so}(n) \to \mathbb{R}^{n\times n}_{sym}$ given by $L(C) = CX - XC$. The adjoint map $L^* : \mathbb{R}^{n\times n}_{sym} \to \mathfrak{so}(n)$ is given by $L^*(\Omega) = \Omega X - X\Omega$. Therefore

$$\{C \in \mathfrak{so}(\mathbb{R}^n) : CX = XC\}^\perp = \{\Omega X - X\Omega : \Omega \in \mathbb{R}^{n\times n}_{sym}\}.$$

Given $V \in T_Y M$ we aim to write $V = BY + YA$ where $A \in \mathfrak{so}(k)$, $B \in \mathfrak{so}(\mathbb{P})$ and furthermore $B = \Omega X - X\Omega$ for $\Omega \in \mathbb{R}^{n\times n}_{sym}$. Under this ansantz, we compute $Y^\top V$, $VY^\top$, and $YV^\top$ and use that $Y^\top Y = I_k$ to find

$$A = Y^\top V, \tag{3.7}$$

$$B = \Omega X - X\Omega = VY^\top - YV^\top - 2YAY^\top. \tag{3.8}$$

Using the formula for $T_Y M$ (3.3) one can check that we actually recover V as $V = BY + YA$ and that $A \in \mathfrak{so}(\mathbb{R}^k)$ and $B \in \mathfrak{so}(\mathbb{P})$, i.e.

$$A + A^\top = 0,\ B + B^\top = 0,\ B1_n = 0. \tag{3.9}$$

3.6 A retraction map

Given $V \in T_Y M$ we aim to find a retraction

$$\mathrm{retr}_Y(V) \in M$$

satisfying (2.5). Write V as $V = BY + YA$ as in (3.7), (3.8). Note we may also see V as $BY + (YAY^\top)Y = (B + (YAY^\top))Y$; this is using the equivalence between multiplication on the right by $O(\mathbb{R}^k)$ and multiplication on the left by $O(\mathrm{im}(X))$, mentioned in Section 3.4. Let $A' = YAY^\top$. Now set

$$\mathrm{retr}_Y(V) = \exp(B)\exp(A')Y. \tag{3.10}$$

The property (2.5) is straightforward given the differential equation satisfied by exp, but it is not as obvious that $\tilde{Y} = \mathrm{retr}_Y(V) \in M$. The condition $\tilde{Y}^\top \tilde{Y} = I$ follows because we are performing left multiplication by orthogonal matrices. To check that $\tilde{Y}\tilde{Y}^\top 1_n = 1_n$ we may compute,

$$\begin{aligned}
&(\exp(B)\exp(A')Y)(\exp(B)\exp(A')Y)^\top 1_n \\
&\quad = \exp(B)\exp(A')YY^\top \exp(-A')\exp(-B)1_n \\
&\quad = \exp(B)YY^\top \exp(-B)1_n \\
&\quad = \exp(B)YY^\top 1_n = \exp(B)1_n = 1_n.
\end{aligned}$$

We have used, successively, that $A'(YY^\top) = (YY^\top)A' = A'$ (so $\exp(-tA')$ commutes with $YY^\top$), that $B1_n = 0$ (so $\exp(-tB)$ fixes 1_n) and that $YY^\top 1_n = 1_n$. Note that the order of the matrix exponentials matters. For example,

$$V \mapsto \exp(A')\exp(B)Y \text{ and } V \mapsto \exp(B)Y\exp(A)$$

are paths satisfying (2.5) but will not lie on the manifold if $A'1 \neq 0$.

3.7 Numerical algorithm

The projection and retraction functions from the previous section allow us to implement gradient descent algorithms in Manopt. In order to tackle the k-means problem, the algorithm we propose entails iteratively solving the manifold optimization relaxation of k-means (3.1), increasing the penalty λ until convergence to a k-means feasible Y. See Algorithm 1.

Algorithm 1: Manifold optimization iteration for k-means clustering

1: $\lambda_0 \leftarrow 0$
2: **repeat**
3: $\quad Y_{n+1} \leftarrow \text{GradientDescent}(f_\lambda)$ // Initialized at Y_n
4: $\quad \lambda_{n+1} \leftarrow 2\lambda_n + 1$
5: **until** $\|Y_-\|_F < \epsilon$

Theorem 6 in [13] guarantees that step 3 in the algorithm finds a stationary point of the objective. The fact that $\lambda \operatorname{Tr}(DYY^\top)$ is bounded for $Y \in M$ suggests the algorithm may converge to a feasible clustering. It would be very interesting to show that Algorithm 1 converges to the actual k-means solution provided a good initialization.

3.8 Numerical simulations

We sample points uniformly from 4 unit balls in $\mathbb{R}^4$ with centers separated by 2.05 (following the stochastic ball model from Definition 1.4.1). We sample 22, 18, 19 and 21 points from each ball respectively. We run Algorithm 1 using Manopt to implement step 3. In Figure 3.1 we plot the

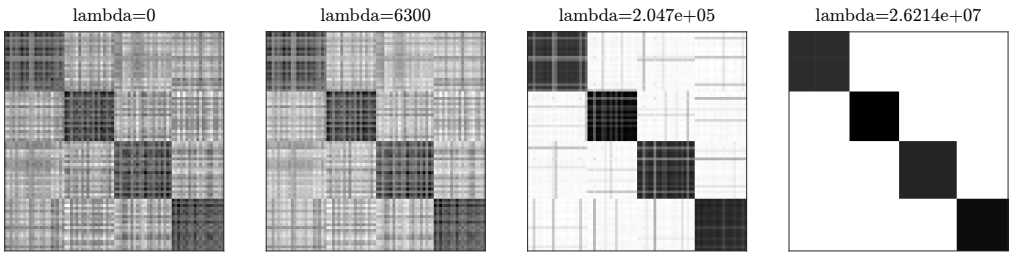


Figure 3.1: Illustration of YY^T for iterations of Algorithm 1 with successive values of λ.
Note the first image is equivalent to a purely spectral method while the last image coincides with the planted solution of the k-means problem. The algorithm is oblivious to the planted order of points. We choose the order where the first points belong to the first cluster, and so on, to simplify visualization.

results.

3.9 Discussion

Before manifold optimization became popular, Burer and Monteiro [20] introduced the idea of using a low rank factorization of a matrix in order to solve a semidefinite program of the form.

$$\begin{aligned} &\text{minimize } \mathrm{Tr}(CX) \\ &\text{subject to } \mathrm{Tr}(A_i X) = b_i \; 1 \leqslant i \leqslant m,\; X \succeq 0 \end{aligned} \tag{3.11}$$

According to the Pataki bound [55], the solution of (3.11) is a matrix $X = YY^\top$ for some $Y \in \mathbb{R}^{n\times p}$ with $\frac{p(p+1)}{2} \leqslant m$. Therefore if we replace the positive semidefinite constraint in (3.11) by $X = YY^\top$ for $Y \in \mathbb{R}^{n\times p}$ the global minimizer of both problems coincide. In their paper, Burer and Monteiro propose an augmented lagrangian iteration in $X = YY^\top$. They prove it converges to a stationary point of their objective. Since the objective is not con-

vex there is a priori no guarantee that it won't converge to some spurious stationary point.

Later work by Journée and collaborators [37] introduced a manifold optimization algorithm and proved that, under somewhat restrictive conditions (not satisfied by our clustering problem), it converges to the global optimizer.

Recent work by Boumal, Voroninski and Bandeira [15] extend Journée's work by showing that the Bourer-Monteiro problem (i.e. the minimization in matrices of the form $YY^\top$) is equivalent to respective SDP for some specific problems. They actually show that for those problems there are no spurious stationary points.

Some natural questions arise: (a) How small can p be chosen with still no spurious stationary point? In their original paper Burer and Monteiro suggested that if the rank of the planted solution is k one should be able to choose $p = k + 1$ or $p = k + 2$. (b) Is it possible to adapt manifold optimization methods to singular manifolds? And in particular, (c) can a theory like this be developed for manifolds with boundary?

To the best of our knowledge the best algorithms that can deal efficiently and reliably with semidefinite programs with non-negative constraints are based on interior point methods [51]. As far as we know, there is no theory that provides convergence guarantees for matrix factorization based algorithms in presence of non-negative constraints; nor even suc-

cessful implementation for algorithms like that for generic SDPs with non-negative constraints.

Chapter 4

Finding the exact solution: tightness in convex optimization

4.1 A semidefinite program relaxation for k-means

Recall Peng and Wei's semidefinite relaxation (k-means sdp) of the k-means problem (k-means).

$$\begin{aligned} \underset{X\in\mathbb{R}^{N\times N}}{\text{minimize}} \quad & \frac{1}{2}\operatorname{trace}(DX) \\ \text{subject to} \quad & X\mathbf{1} = \mathbf{1},\ \operatorname{trace}(X) = k,\ X \geqslant 0,\ X \succeq 0. \end{aligned} \tag{k-means sdp}$$

In this section we show (k-means sdp) is typically tight under the stochastic ball model. We do it in two different ways, one appears in [8] and the other one in [34]. The basic idea is (i) to find a deterministic condition on the set

This chapter is based on two publications:
Pranjal Awasthi, Afonso S. Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, Rachel Ward. *Relax, no need to round: Integrality of clustering formulations.* Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, pp. 191-200. ACM, 2015.
Takayuki Iguchi, Dustin G. Mixon, Jesse Peterson, Soledad Villar. *Probably certifiably correct* k*-means clustering* Mathematical Programming, 2016 (to appear).
In both of the papers the author contributed in developing the main ideas of the paper, the mathematical proofs and numerical experiments.

of points under which the relaxation finds the k-means-optimal solution, and (ii) to discuss when this deterministic condition is satisfied with high probability under the stochastic ball model.

To find the dual program to (k-means sdp) we leverage the cone programming theory from Section 2.1.1. In our case, $c = -D$, $x = X$, and $\mathcal{K}$ is simply the cone of positive semidefinite matrices (as is $\mathcal{K}^*$). Before we determine $\mathcal{L}$, we need to interpret the remaining constraints in (k-means sdp). To this end, we note that $\mathrm{Tr}(X) = k$ is equivalent to $\langle X, I \rangle = k$, $X1 = 1$ is equivalent to having

$$\left\langle X, \frac{1}{2}(e_i 1^\top + 1 e_i^\top) \right\rangle = 1 \qquad \forall i \in \{1, \dots, N\},$$

and $X \succcurlyeq 0$ is equivalent to having

$$\left\langle X, \frac{1}{2}(e_i e_j^\top + e_j e_i^\top) \right\rangle \geqslant 0 \qquad \forall i, j \in \{1, \dots, N\},\ i \leqslant j.$$

(These last two equivalences exploit the fact that X is symmetric.) As such, we can express the remaining constraints in (k-means sdp) using a linear operator $\mathcal{A}$ that sends any matrix X to its inner products with I, $\{\frac{1}{2}(e_i 1^\top + 1 e_i^\top)\}_{i=1}^N$, and $\{\frac{1}{2}(e_i e_j^\top + e_j e_i^\top)\}_{i,j=1, i\leqslant j}^N$. Note that the remaining constraints in (k-means sdp) are equivalent to having $b - \mathcal{A}x \in \mathcal{L}$, where $b = k \oplus 1 \oplus 0$ and $\mathcal{L} = 0 \oplus 0 \oplus \mathbb{R}^{N(N+1)/2}_{\succcurlyeq 0}$. Writing $y = z \oplus \alpha \oplus (-\beta)$, the dual of (k-means sdp) is then given by

$$\underset{z\in\mathbb{R},\,\alpha\in\mathbb{R}^N,\,\beta\in\mathbb{R}^{N\times N}}{\text{minimize}} \quad kz+\sum_{i=1}^{N}\alpha_i \qquad \text{(k-means sdp dual)}$$

$$\text{subject to } Q := zI+\sum_{i=1}^{N}\alpha_i\cdot\frac{1}{2}(e_i1^\top+1e_i^\top)-\sum_{i=1}^{N}\sum_{j=i}^{N}\beta_{ij}\cdot\frac{1}{2}(e_ie_j^\top+e_je_i^\top)+D\succeq 0$$

$$\beta\geqslant 0$$

For notational simplicity, from this point forward, we organize indices according to clusters. For example, 1_a shall denote the indicator function of the ath cluster. Also, we shuffle the rows and columns of X and D into blocks that correspond to clusters; for example, the (i,j)th entry of the (a,b)th block of D is given by $D_{ij}^{(a,b)}$. We also index α in terms of clusters; for example, the ith entry of the ath block of α is denoted $\alpha_{a,i}$. For β, we identify

$$\beta := \sum_{i=1}^{N}\sum_{j=i}^{N}\beta_{ij}\cdot\frac{1}{2}(e_ie_j^\top+e_je_i^\top).$$

Indeed, when $i\leqslant j$, the (i,j)th entry of β is β_{ij}. We also consider β as having its rows and columns shuffled according to clusters, so that the (i,j)th entry of the (a,b)th block is $\beta_{ij}^{(a,b)}$.

With this notation, the following proposition characterizes all possible dual certificates of (k-means sdp):

Proposition 4.1.1. *Take* $X := \sum_{a=1}^{k}\frac{1}{n_a}1_a1_a^\top$, *where* n_a *denotes the number of points in cluster* a. *The following are equivalent:*

(a) X *is a solution to the semidefinite relaxation* (k-means sdp).

(b) Every solution to the dual program (k-means sdp dual) *satisfies*

$$Q^{(a,a)}1 = 0, \qquad \beta^{(a,a)} = 0 \qquad \forall a \in \{1,\dots,k\}.$$

(c) Every solution to the dual program (k-means sdp dual) *satisfies*

$$\alpha_{a,r} = -\frac{1}{n_a}z + \frac{1}{n_a^2}1^\top D^{(a,a)}1 - \frac{2}{n_a}e_r^\top D^{(a,a)}1 \qquad \forall a \in \{1,\dots,k\},\ r \in a.$$

Proof. (a)⇔(b): By complementary slackness, (a) is equivalent to having both

$$\langle A^*y - c, X\rangle = 0 \tag{4.1}$$

and

$$\langle y, b - A(X)\rangle = 0. \tag{4.2}$$

Since $Q \succeq 0$, we have

$$\langle A^*y - c, X\rangle = \langle Q, X\rangle = \left\langle Q, \sum_{t=1}^{k}\frac{1}{n_t}1_t1_t^\top\right\rangle = \sum_{t=1}^{k}\frac{1}{n_t}1_t^\top Q1_t \geqslant 0,$$

with equality if and only if $Q1_a = 0$ for every $a \in \{1,\dots,k\}$. Next, we recall that $y = z \oplus \alpha \oplus (-\beta)$, $b - A(X) \in L = 0 \oplus 0 \oplus \mathbb{R}^{N(N+1)/2}_{\geqslant 0}$, and $b = k \oplus 1 \oplus 0$. As such, (4.2) is equivalent to β having disjoint support with $\{\langle X, \frac{1}{2}(e_ie_j^\top + e_je_i^\top)\rangle\}_{i,j=1,i\leqslant j}^{N}$, i.e., $\beta^{(a,a)} = 0$ for every cluster a.

(b)⇒(c): Take any solution to the dual SDP (k-means sdp dual), and note that

$$\begin{aligned} Q^{(a,a)} &= zI + \left(\sum_{t=1}^{k}\sum_{i\in t}\alpha_{t,i}\cdot\frac{1}{2}(e_{t,i}1^\top + 1e_{t,i}^\top)\right)^{(a,a)} - \beta^{(a,a)} + D^{(a,a)} \\ &= zI + \sum_{i\in a}\alpha_{a,i}\cdot\frac{1}{2}(e_i1^\top + 1e_i^\top) + D^{(a,a)}, \end{aligned} \tag{4.3}$$

where the 1 vectors in the second line are n_a-dimensional (instead of N-dimensional, as in the first line), and similarly for e_i (instead of $e_{t,i}$). We now consider each entry of $Q^{(a,a)}1$, which is zero by assumption:

$$\begin{aligned}
0 &= e_r^\top Q^{(a,a)} 1 \\
&= e_r^\top \left(zI + \sum_{i \in a} \alpha_{a,i} \cdot \frac{1}{2}(e_i 1^\top + 1 e_i^\top) + D^{(a,a)} \right) 1 \\
&= z + \sum_{i \in a} \alpha_{a,i} \cdot \frac{1}{2}(e_r^\top e_i 1^\top 1 + e_r^\top 1 e_i^\top 1) + e_r^\top D^{(a,a)} 1 \\
&= z + \sum_{i \in a} \alpha_{a,i} \cdot \frac{1}{2}(n_a \delta_{ir} + 1) + e_r^\top D^{(a,a)} 1. \qquad (4.4)
\end{aligned}$$

As one might expect, these n_a linear equations determine the variables $\{\alpha_{a,i}\}_{i \in a}$. To solve this system, we first observe

$$\begin{aligned}
0 &= 1^\top Q^{(a,a)} 1 \\
&= 1^\top \left(zI + \sum_{i \in a} \alpha_{a,i} \cdot \frac{1}{2}(e_i 1^\top + 1 e_i^\top) + D^{(a,a)} \right) 1 \\
&= n_a z + \sum_{i \in a} \alpha_{a,i} \cdot \frac{1}{2}(1^\top e_i 1^\top 1 + 1^\top 1 e_i^\top 1) + 1^\top D^{(a,a)} 1 \\
&= n_a z + n_a \sum_{i \in a} \alpha_{a,i} + 1^\top D^{(a,a)} 1,
\end{aligned}$$

and so rearranging gives

$$\sum_{i \in a} \alpha_{a,i} = -z - \frac{1}{n_a} 1^\top D^{(a,a)} 1.$$

We use this identity to continue (4.4):

$$\begin{aligned}
0 &= z + \sum_{i \in a} \alpha_{a,i} \cdot \frac{1}{2}(n_a \delta_{ir} + 1) + e_r^\top D^{(a,a)} 1 \\
&= z + \frac{n_a}{2} \alpha_{a,r} + \frac{1}{2} \sum_{i \in a} \alpha_{a,i} + e_r^\top D^{(a,a)} 1 \\
&= z + \frac{n_a}{2} \alpha_{a,r} + \frac{1}{2} \left(-z - \frac{1}{n_a} 1^\top D^{(a,a)} 1 \right) + e_r^\top D^{(a,a)} 1,
\end{aligned}$$

and rearranging yields the desired formula for $\alpha_{a,r}$.

(c)⇒(a): Take any solution to the dual SDP (k-means sdp dual). Then by assumption, the dual objective at this point is given by

$$\begin{aligned}
kz + \sum_{t=1}^{k} \sum_{i \in t} \alpha_{t,i} &= kz + \sum_{t=1}^{k} \sum_{i \in t} \left(-\frac{1}{n_t} z + \frac{1}{n_t^2} 1^\top D^{(t,t)} 1 - \frac{2}{n_t} e_i^\top D^{(t,t)} 1 \right) \\
&= -\sum_{t=1}^{k} \frac{1}{n_t} 1^\top D^{(t,t)} 1 \\
&= -\operatorname{Tr}(DX),
\end{aligned}$$

i.e., the primal objective (k-means sdp) evaluated at X. Since X is feasible in the primal SDP, we conclude that X is optimal by strong duality. □

Remark 4.1.1 (Pointed out by Xiaodong Li on our preprint [33]). The statement $Q^{(a,a)} 1 = 0$ implies $Q1 = 0$.

Proof. Let $a \in \{1, \ldots, k\}$ and let R be a $N \times N$ symmetric positive semidefinite matrix with blocks $R^{(a,a)} = 1_a 1_a^\top$, $R^{(b,b)} = I_b$, $R^{(b,a)} = 0$ for all $b \neq 0$. Then $L := R^\top Q R$ is a symmetric positive semidefinite matrix where $L^{(a,a)} = 0$, therefore for every (a, b) we have $L^{(b,a)} = 0$, but note that $L^{(b,a)} = Q^{(b,a)} 1_a 1_a^\top$. □

The following subsection will leverage Proposition 4.1.1 to identify a condition on D that implies that the SDP (k-means sdp) relaxation is tight.

4.1.1 Dual certificate from separation condition

Based on numerical observations we now present a *guess* for the matrix Q that satisfies the required constraints. We set for all $a \neq b$

$$e_r^\top Q^{(a,b)} e_s = \frac{1}{n} e_r^\top D^{(a,b)} 1 + \frac{1}{n} 1^\top D^{(a,b)} e_s - e_r^\top D^{(a,b)} e_s - \frac{1}{n^2} 1^\top D^{(a,b)} 1. \quad (4.5)$$

Observe that the above definition essentially combines (for two points r, s in clusters a, b respectively) (i) the average distance of r to the cluster b, the average distance of s to cluster a, the distance between r and s, and the average distance between the two clusters.

Note that $Q^{(a,b)} 1 = 0$ and $Q^{(b,a)} 1 = Q(a,b)^\top 1 = 0$. By definition of Q we will require for all r, s, that

$$\begin{aligned} e_r^\top Q^{(a,b)} e_s &= \frac{1}{n} e_r^\top D^{(a,b)} 1 + \frac{1}{n} 1^\top D^{(a,b)} e_s - D_{rs}^{(a,b)} - \frac{1}{n^2} 1^\top D^{(a,b)} 1 \\ &= -z\frac{1}{n} + \frac{1}{2n}\left[\left(\frac{1}{n} 1^\top D^{(a,a)} 1 - 2 e_r^\top D^{(a,a)} 1\right) + \left(\frac{1}{n} 1^\top D^{(b,b)} 1 - 2 e_s^\top D^{(b,b)} 1\right)\right] \\ &\quad - \frac{1}{2}\beta_{rs}^{(a,b)} + D_{rs}^{(a,b)}. \end{aligned} \quad (4.6)$$

This is satisfied for non-negative β's precisely when

$$\begin{aligned} &2D_{rs}^{(a,b)} - \frac{1}{n} e_r^\top D^{(a,b)} 1 - \frac{1}{n} 1^\top D^{(a,b)} e_s + \frac{1}{n^2} 1^\top D^{(a,b)} 1 \quad (4.7) \\ &\geqslant \frac{e_r^\top D^{(a,a)} 1}{n} + \frac{e_s^\top D^{(b,b)} 1}{n} - \frac{1}{2}\left(\frac{1^\top D^{(a,a)} 1}{n^2} + \frac{1^\top D^{(b,b)} 1}{n^2}\right) + \frac{1}{n} z, \ \forall_{a \neq b} \forall_{r,s}. \end{aligned}$$

It remains to ensure that $Q \succeq 0$. By construction, $Q^{(a,b)}1 = 0 \quad \forall_{a,b}$ so we just need to ensure that, for all x perpendicular to the subspace Λ spanned by $\{1^{(a)}\}_{a=1}^{k}$ that

$$x^\top Q x > 0. \tag{4.8}$$

Since in particular $x \perp 1$, as a consequence of (4.3) and (4.5) the expression greatly simplifies to:

$$z x^\top x + 2x^\top (\sum_a D^{(a,a)})x - x^\top D x > 0, \tag{4.9}$$

which means that we simply need

$$z > \frac{x^\top D x}{x^\top x} - \frac{2x^\top (\sum_a D^{(a,a)})x}{x^\top x}, \quad \forall_{x \perp \Lambda}. \tag{4.10}$$

Now, we can decompose the squared euclidean distance matrix D as

$$D = \nu 1^\top - 2\Phi^\top \Phi + 1\nu^\top,$$

where ν is the $N \times 1$ vector whose (a,i)th entry is $\|x_{a,i}\|_2^2$, and Φ is the $m \times N$ matrix whose (a,i)th column is $x_{a,i}$. Since $\Phi^\top \Phi$ is positive semidefinite then (4.10) can be stated as

$$z > 4 \max_a \max_{x \perp 1} \frac{x^\top \Phi^{(a)\top} \Phi^{(a)} x}{x^\top x} \tag{4.11}$$

Since we need the existence of a z to satisfy both (4.11) and (4.7) we need that $\forall_{a \neq b} \forall_{r,s}$,

$$\begin{aligned} 2D_{rs}^{(a,b)} - \frac{1}{n} e_r^\top D^{(a,b)} 1 - \frac{1}{n} 1^\top D^{(a,b)} e_s + \frac{1}{n^2} 1^\top D^{(a,b)} 1 > \frac{e_r^\top D^{(a,a)} 1}{n} + \frac{e_s^\top D^{(b,b)} 1}{n} \\ - \frac{1}{2}\left(\frac{1^\top D^{(a,a)} 1}{n^2} + \frac{1^\top D^{(b,b)} 1}{n^2}\right) + \frac{1}{n}\left(4 \max_a \max_{x \perp 1} \left| \frac{x^\top \Phi^{(a)\top} \Phi^{(a)} x}{x^\top x} \right| \right) \end{aligned} \tag{4.12}$$

This gives us the main Lemma of this section:

Lemma 4.1.2. *If, for all clusters* $a \neq b$ *and for all indices* r, s *we have*

$$2D_{rs}^{(a,b)} - \frac{1}{n} e_r^\top D^{(a,b)} 1 - \frac{1}{n} 1^\top D^{(a,b)} e_s + \frac{1}{n^2} 1^\top D^{(a,b)} 1 > \frac{e_r^\top D^{(a,a)} 1}{n} + \frac{e_s^\top D^{(b,b)} 1}{n} - \frac{1}{2}\left(\frac{1^\top D^{(a,a)} 1}{n^2} + \frac{1^\top D^{(b,b)} 1}{n^2}\right) + \frac{1}{n}\left(4 \max_a \max_{x \perp 1} \left|\frac{x^\top \Phi^{(a)\top} \Phi^{(a)} x}{x^\top x}\right|\right) \quad (4.13)$$

then the k-means SDP has a unique solution and it coincides with the intended cluster solution.

Definition 4.1.1 (Average separation condition)**.** For cluster c define $x_c = \sum_{y \in c} y$ the mean of the cluster. A clustering instance satisfies average separation if, for all clusters $a \neq b$ and for all indices r, s we have

$$2\|x_r - x_s\|^2 - \|x_r - x_b\|^2 - \|x_s - x_a\|^2 - \|x_r - x_a\|^2 - \|x_s - x_b\|^2 + \|x_a - x_b\|^2 > \frac{1}{n}\left(4 \max_a \max_{x \perp 1} \left|\frac{x^\top \Phi^{(a)\top} \Phi^{(a)} x}{x^\top x}\right|\right) \quad (4.14)$$

Note that (4.13) and (4.14) are equivalent due the parallelogram identity. Hence, we have proved the following theorem.

Theorem 4.1.3. *If a euclidean clustering instance satisfies the average separation condition from Definition 4.1.1, then the corresponding* k*-means SDP for the instance has unique integral solution equal to the* k*-means optimal solution, and corresponding to this clustering.*

Now we state that under the stochastic ball model, under separation of at least $2\sqrt{2(1+1/m)}$, average separation is satisfied for large enough n. We obtain the following theorem:

Theorem 4.1.4. *Under the $(\mathcal{D},\gamma,n)$-stochastic ball model in $\mathbb{R}^m$, if*

$$\Delta > \sqrt{8\left(1+\frac{3}{\log n}+\frac{1}{(\log n)^2}\right)\frac{\theta}{m}+8}$$

where $\theta = \mathbb{E}(\|x_{a,r}-\gamma_a\|^2) < 1$. There is a universal constant $c>0$ such that with probability exceeding $1-4mk\exp\left(\frac{-cn}{(\log n)^2 m\Delta^2}\right)$ the k-means SDP has a unique integral solution which coincides with the intended cluster solution.

Remark 4.1.2. In the limit $n\to\infty$, the probability of success goes to 1 and the separation distance goes to $2\sqrt{2(1+\frac{\theta}{m})}$.

In the rest of this section we prove Theorem 4.1.4.

Lemma 4.1.5. *Under the same hypothesis as before, then the LHS of* (4.14) *for fixed a,b, with probability at least $1-m\exp(-cn\epsilon^2/\Delta^2)$ (with c an absolute constant) has minimum at least*

$$\begin{cases}\Delta^2/2-4-\epsilon & \text{if } \Delta\leqslant 4\\ (\Delta-2)^2-\epsilon & \text{if } \Delta>4\end{cases} \tag{4.15}$$

In particular it is positive for center separation $\Delta > 2\sqrt{2}$ with high probability.

Proof. Without loss of generality assume $\gamma_a=(0,\dots,0)\in\mathbb{R}^m$ and $\gamma_b=(\Delta,0,\dots,0)\in\mathbb{R}^m$. First we search for

$$\begin{aligned}\min\quad & 2\|x_r-x_s\|^2-\|x_r-\gamma_b\|^2-\|x_s-\gamma_a\|^2 \\ & \qquad -\|x_r-\gamma_a\|^2-\|x_s-\gamma_b\|^2+\|x_a-\gamma_b\|^2 \\ \text{subject to}\quad & \|x_r-\gamma_a\|^2\leqslant 1,\quad \|x_s-\gamma_b\|^2\leqslant 1\end{aligned} \tag{4.16}$$

This is a calculus exercise one can solve using Lagrange multipliers. If $\Delta \leqslant 4$ then the minimum is attained at points such that $x_{r(1)} = \Delta/4$, $x_{s(1)} = \Delta/2$ and $x_{r(i)} = x_{s(i)}$ for all $i = 2, \ldots, m$. When $\Delta > 4$ the minimum is attained in $x_r = (1, 0, \ldots, 0)$, $x_s = (\Delta - 1, 0, \ldots, 0)$.

Let $r_{a,i} \sim x_{a,i} - \gamma a$. Now expanding the squares and using Cauchy-Schwarz we observe that the difference between (4.16) and the LHS of (4.14) in absolute value is at most

$$2\left\|\frac{1}{n}\sum_{i=1}^{n} r_{a,i}\right\|^2 + 2\left\|\frac{1}{n}\sum_{i=1}^{n} r_{b,i}\right\|^2 + C\Delta\left\|\frac{1}{n}\sum_{i=1}^{n} r_{a,i}\right\| + C\Delta\left\|\frac{1}{n}\sum_{i=1}^{n} r_{b,i}\right\|$$

with C an absolute constant.

Since $\mathbb{E}r = 0$ and $\|r\|_2^2 \leqslant 1$ almost surely, one may lift

$$X_{a,i} := \begin{bmatrix} 0 & r_{a,i}^\top \\ r_{a,i} & 0 \end{bmatrix}$$

and apply the Matrix Hoeffding inequality [61] to conclude that

$$\Pr\left(\left\|\sum_{i=1}^{n} r_{a,i}\right\|_2 \geqslant t\right) \leqslant me^{-t^2/8n}.$$

Taking $\epsilon = \frac{cn}{\Delta}\epsilon$ gives us the result. □

Proof of Theorem 4.1.4. We bound the RHS of (4.14). Given our distributional model, we can then write $\Phi = \widetilde{\Phi} + C$ where $\widetilde{\Phi}$ has independent and identically distributed columns drawn from μ, and C is a rank k matrix whose columns are constant within any cluster: the $((a,r),(b,s))$th column is the shift $x_b - x_a$, and the $((a,r),(a,s))$th column is zero.

Recall that Λ is the k-dimensional subspace spanned by $\{1^{(a)}\}_{a=1}^{k}$. Since $C^\top z = 0$ for $z \perp \Lambda$ we have,

$$\frac{1}{n}\left[4\max_{z\perp\Lambda}\frac{z^\top \Phi^{(a)\top}\Phi^{(a)}z}{z^\top z}\right] = \frac{1}{n}\left[4\max_{z\perp\Lambda}\frac{z^\top \widetilde{\Phi}^{(a)\top}\widetilde{\Phi}^{(a)}z}{z^\top z}\right] \leqslant \frac{4}{n}\sigma_{\max}(\widetilde{\Phi}^{(a)})^2.$$

The columns of $\widetilde{\Phi}$ are the centered points, $\tilde{x}_{a,r}$. Let θ be the expected value $\theta = \mathbb{E}(\|\tilde{x}_{a,r}\|^2)$. The columns of $\sqrt{\frac{m}{\theta}}\widetilde{\Phi}$ are independent isotropic random vectors and $\|\sqrt{\frac{m}{\theta}}\tilde{x}_{a,r}\|_2 \leqslant \sqrt{m/\theta}$. We use quantitative bounds on the spectra of such matrices. By Theorem 5.41 of [64], we have that for every $t \geqslant 0$,

$$\mathbb{P}\left[\sigma_{\max}\left(\sqrt{\frac{m}{\theta}}\widetilde{\Phi}^{(a)}\right) > \sqrt{n} + t\sqrt{\frac{m}{\theta}}\right] \leqslant 2m\exp(-ct^2), \tag{4.17}$$

where $c > 0$ is an absolute constant. Taking $t = s\sqrt{\frac{n\theta}{m}}$, we find that $\frac{4}{n}\sigma_{\max}(\widetilde{\Phi}^{(a)})^2 \leqslant 4\theta(1+s)^2\frac{1}{m}$ with probability at least $1 - 2m\exp(-cns^2/m)$.

By a union bound, we have that

$$\frac{1}{n}\left(4\max_{a}\max_{z\perp 1}\frac{z^\top \Phi^{(a)\top}\Phi^{(a)}z}{z^\top z}\right) \leqslant 4\theta(1+s)^2\frac{1}{m} \tag{4.18}$$

with probability exceeding $1 - 2mk\exp(-cns^2/m)$

Combining the bound of the LHS in (4.14) from Lemma 4.1.5 with (4.18) we obtain that the sufficient condition for integrality of the k-means SDP is satisfied with probability exceeding $1 - 2mk\exp(-cns^2/m) - mk\exp(-c'n\epsilon^2/\Delta^2)$ if

$$\frac{\Delta^2}{2} - 4 - \epsilon > 4(1+s)^2\frac{\theta}{m}$$

which holds once the centers of the clusters are separated by euclidean distance $\Delta > \sqrt{8(1+s)^2\frac{\theta}{m} + \epsilon + 8}$. Fixing the parameter $s = \frac{1}{\log n}$ and $\epsilon = \frac{8\theta}{m\log n}$ the above analysis proves Theorem 4.1.4. □

4.1.2 Dual certificate from spectral condition

In this section we derive a completely different dual certificate for the k-means SDP. Recall Proposition 4.1.1, who characterizes acceptable dual certificates (z, α, β) but unfortunately fails to uniquely determine a certificate. In the previous section we presented a dual certificate based on a separation condition. In this subsection, we will motivate the application of additional constraints on dual certificates so as to identify certifiable instances.

We start by reviewing the characterization of dual certificates (z, α, β) provided in Proposition 4.1.1. In particular, α is completely determined by z, and so z and β are the only remaining free variables. Indeed, for every $a, b \in \{1, \ldots, k\}$, we have

$$\begin{aligned}
&\left(\sum_{t=1}^{k} \sum_{i \in t} \alpha_{t,i} \cdot \frac{1}{2}(e_{t,i} 1^\top + 1 e_{t,i}^\top) \right)^{(a,b)} \\
&\quad = \sum_{i \in a} \alpha_{a,i} \cdot \frac{1}{2} e_i 1^\top + \sum_{j \in b} \alpha_{b,j} \cdot \frac{1}{2} 1 e_j^\top \\
&\quad = -\frac{1}{2}\left(\frac{1}{n_a} + \frac{1}{n_b}\right) z + \sum_{i \in a} \left(\frac{1}{n_a^2} 1^\top D^{(a,a)} 1 - \frac{2}{n_a} e_i^\top D^{(a,a)} 1 \right) \frac{1}{2} e_i 1^\top \\
&\qquad + \sum_{j \in b} \left(\frac{1}{n_b^2} 1^\top D^{(b,b)} 1 - \frac{2}{n_b} e_j^\top D^{(b,b)} 1 \right) \frac{1}{2} 1 e_j^\top,
\end{aligned}$$

and so since

$$Q = zI + \sum_{t=1}^{k} \sum_{i \in t} \alpha_{t,i} \cdot \frac{1}{2}(e_{t,i} 1^\top + 1 e_{t,i}^\top) - \frac{1}{2}\beta + D,$$

we may write $Q = z(I - E) + M - B$, where

$$E^{(a,b)} := \frac{1}{2}\left(\frac{1}{n_a} + \frac{1}{n_b}\right)11^\top \tag{4.19}$$

$$M^{(a,b)} := D^{(a,b)} + \sum_{i\in a}\left(\frac{1}{n_a^2}1^\top D^{(a,a)}1 - \frac{2}{n_a}e_i^\top D^{(a,a)}1\right)\frac{1}{2}e_i1^\top + \sum_{j\in b}\left(\frac{1}{n_b^2}1^\top D^{(b,b)}1 - \frac{2}{n_b}e_j^\top D^{(b,b)}1\right)\frac{1}{2}1e_j^\top \tag{4.20}$$

$$B^{(a,b)} = \frac{1}{2}\beta^{(a,b)}$$

for every $a, b \in \{1,\dots,k\}$. The following is one way to formulate our task: Given D and a clustering X (which in turn determines E and M), determine whether there exist feasible z and B such that $Q \succeq 0$; here, feasibility only requires B to be symmetric with nonnegative entries and $B^{(a,a)} = 0$ for every $a \in \{1,\dots,k\}$. We opt for a slightly more modest goal: Find $z = z(D,X)$ and $B = B(D,X)$ such that $Q \succeq 0$ for a large family of D's.

Before determining z and B, we first analyze E:

Lemma 4.1.6. *Let* E *be the matrix defined by* (4.19)*. Then* $\operatorname{rank}(E) \in \{1,2\}$*. The eigenvalue of largest magnitude is* $\lambda \geqslant k$*, and when* $\operatorname{rank}(E) = 2$*, the other nonzero eigenvalue of* E *is negative. The eigenvectors corresponding to nonzero eigenvalues lie in the span of* $\{1_a\}_{a=1}^k$*.*

Proof. Writing

$$E = \sum_{a=1}^k\sum_{b=1}^k \frac{1}{2}\left(\frac{1}{n_a} + \frac{1}{n_b}\right)1_a1_b^\top = \frac{1}{2}\left(\sum_{a=1}^k \frac{1}{n_a}1_a\right)1^\top + \frac{1}{2}1\left(\sum_{b=1}^k \frac{1}{n_b}1_b\right)^\top,$$

we see that $\text{rank}(E) \in \{1, 2\}$, and it is easy to calculate $1^\top E 1 = Nk$ and $\text{Tr}(E) = k$. Observe that

$$\lambda = \sup_{\substack{x \in \mathbb{R}^N \\ \|x\|_2 = 1}} x^\top E x \geqslant \frac{1}{N} 1^\top E 1 = k,$$

and combining with $\text{rank}(E) \leqslant 2$ and $\text{Tr}(E) = k$ then implies that the other nonzero eigenvalue (if there is one) is negative. Finally, any eigenvector of E with a nonzero eigenvalue necessarily lies in the column space of E, which is a subspace of $\text{span}\{1_a\}_{a=1}^k$ by the definition of E. □

When finding z and B such that $Q = z(I - E) + M - B \succeq 0$, it will be useful that $I - E$ has only one negative eigenvalue. Let v_0 denote the corresponding eigenvector. Then combining Lemma 4.1.6 and Remark 4.1.1 we know v_0 is also an eigenvector of $M - B$. Since

$$\begin{aligned} 0 &= (Q 1_b)_a \\ &= \Big((z(I - E) + M - B) 1_b \Big)_a \\ &= -z E^{(a,b)} 1 + M^{(a,b)} 1 - B^{(a,b)} 1 \\ &= -z \frac{n_a + n_b}{2n_a} 1 + M^{(a,b)} 1 - B^{(a,b)} 1, \end{aligned} \tag{4.21}$$

then, in order for there to exist a vector $B^{(a,b)} 1 \geqslant 0$ that satisfies (4.21), z must satisfy

$$z \frac{n_a + n_b}{2n_a} \leqslant \min(M^{(a,b)} 1),$$

and since z is independent of (a, b), we conclude that

$$z \leqslant \min_{\substack{a,b \in \{1,\ldots,k\} \\ a \neq b}} \frac{2n_a}{n_a + n_b} \min(M^{(a,b)} 1). \tag{4.22}$$

Now it is time to make a choice for the dual certificate. In order to ensure $z(I - E) + M - B \succeq 0$ for as many instances of D as possible, we intend to choose z as large as possible. We choose B which satisfies (4.21) for every (a, b), even when z satisfies equality in (4.22). Indeed, we define

$$u_{(a,b)} := M^{(a,b)}1 - z\frac{n_a + n_b}{2n_a}1, \qquad \rho_{(a,b)} := u_{(a,b)}^\top 1,$$

$$B^{(a,b)} := \frac{1}{\rho_{(b,a)}} u_{(a,b)} u_{(b,a)}^\top \tag{4.23}$$

for every $a, b \in \{1, \dots, k\}$ with $a \neq b$. Then by design, B immediately satisfies (4.21). Also, note that $\rho_{(a,b)} = \rho_{(b,a)}$, and so $B^{(b,a)} = (B^{(a,b)})^\top$, meaning B is symmetric. Finally, we necessarily have $u_{(a,b)} \geqslant 0$ (and thus $\rho_{(a,b)} \geqslant 0$) by (4.22), and we implicitly require $\rho_{(a,b)} > 0$ for division to be permissible. As such, we also have $B^{(a,b)} \geqslant 0$, as desired.

Now that we have selected z and B, it remains to check that $Q \succeq 0$. By construction, we already have $\Lambda = \operatorname{span}\{1_a\}_{a=1}^k$ in the nullspace of Q, and so it suffices to ensure

$$0 \preceq P_{\Lambda^\perp} Q P_{\Lambda^\perp} = P_{\Lambda^\perp}\Big(z(I - E) + M - B\Big)P_{\Lambda^\perp} = zP_{\Lambda^\perp} + P_{\Lambda^\perp}(M - B)P_{\Lambda^\perp}.$$

Here, $P_{\Lambda^\perp}$ denotes the orthogonal projection onto the orthogonal complement of Λ. Rearranging then gives the following result:

Theorem 4.1.7. *Take* $X := \sum_{t=1}^k \frac{1}{n_t} 1_t 1_t^\top$, *where* n_t *denotes the number of points in cluster* t. *Consider* M *defined by* (4.20), *pick* z *so as to satisfy equality in* (4.22), *take* B *defined by* (4.23), *and let* Λ *denote the span of* $\{1_t\}_{t=1}^k$. *Then* X *is a solution*

to the semidefinite relaxation (k-means sdp) *if*

$$P_{\Lambda^\perp}(B - M)P_{\Lambda^\perp} \preceq zP_{\Lambda^\perp}. \tag{4.24}$$

The next subsection leverages this sufficient condition to establish that the Peng–Wei SDP (k-means sdp) is typically tight under the stochastic ball model.

4.1.3 Integrality of the relaxation under the stochastic ball model

We first note that our sufficient condition (4.24) is implied by

$$\|P_{\Lambda^\perp}MP_{\Lambda^\perp}\| + \|P_{\Lambda^\perp}BP_{\Lambda^\perp}\| \leqslant z$$

since $P_{\Lambda^\perp}|_{\Lambda^\perp} = zI_{\Lambda^\perp}$ and $\Lambda \subset \ker(P_{\Lambda^\perp}(B-M)P_{\Lambda^\perp})$. By further analyzing the left-hand side above (see Section 4.1.3.1), we arrive at the following corollary:

Corollary 4.1.8. *Take* $X := \sum_{t=1}^{k} \frac{1}{n_t} 1_t 1_t^\top$, *where* n_t *denotes the number of points in cluster* t. *Let* Ψ *denote the* $m \times N$ *matrix whose* (a,i)*th column is* $x_{a,i} - c_a$, *where*

$$c_a := \frac{1}{n_a} \sum_{i \in a} x_{a,i}$$

denotes the empirical center of cluster a. *Consider* M *defined by* (4.20), *pick* z *so as to satisfy equality in* (4.22), *and take* $\rho_{(a,b)}$ *defined by* (4.23). *Then* X *is a solution to the semidefinite relaxation* (k-means sdp) *if*

$$2\|\Psi\|^2 + \sum_{a=1}^{k} \sum_{b=a+1}^{k} \frac{\|P_{1^\perp}M^{(a,b)}1\|_2 \|P_{1^\perp}M^{(b,a)}1\|_2}{\rho_{(a,b)}} \leqslant z.$$

In Section 4.1.3.2, we leverage the stochastic ball model to bound each term in Corollary 4.1.8, and in doing so, we identify a regime in which the data points typically satisfy the sufficient condition given in Corollary 4.1.8:

Theorem 4.1.9. *The* k*-means semidefinite relaxation* (k-means sdp) *recovers the planted clusters in the* $(\mathcal{D}, \gamma, n)$*-stochastic ball model with probability* $1 - e^{-\Omega_{\mathcal{D},\gamma}(n)}$ *provided* $\Delta > 2 + k^2/m$.

When $k = o(m^{1/2})$, Theorem 4.1.9 is near-optimal, and in this sense, it's a significant improvement over the sufficient condition in the previous section

$$\Delta > 2\sqrt{2(1 + \frac{1}{m})} \tag{4.25}$$

given in [8]. However, there are regimes (e.g., $k = m$) for which (4.25) is much better, leaving open the question of what the optimal bound is.

4.1.3.1 Proof of Corollary 4.1.8

It suffices to have

$$\|P_{A^\perp} M P_{A^\perp}\| + \|P_{A^\perp} B P_{A^\perp}\| \leqslant z. \tag{4.26}$$

We will bound the terms in (4.26) separately and then combine the bounds to derive a sufficient condition for Theorem 4.1.7. To bound the first term in (4.26), recall $D = \nu 1^\top - 2\Phi^\top \Phi + 1\nu^\top$ were ν is the $N \times 1$ vector whose (a,i)th entry is $\|x_{a,i}\|_2^2$, and Φ is the $m \times N$ matrix whose (a,i)th column is

$x_{a,i}$. With this, we appeal to the blockwise definition of M (4.20):

$$\begin{aligned}\|P_{\Lambda^\perp} M P_{\Lambda^\perp}\| = \|P_{\Lambda^\perp} D P_{\Lambda^\perp}\| &= \|P_{\Lambda^\perp}(\nu 1^\top - 2\Phi^\top\Phi + 1\nu^\top)P_{\Lambda^\perp}\| \\ &= 2\|P_{\Lambda^\perp}\Phi^\top\Phi P_{\Lambda^\perp}\| = 2\|\Phi P_{\Lambda^\perp}\|^2 = 2\|\Psi\|^2.\end{aligned}$$

For the second term in (4.26), we first write the decomposition

$$B = \sum_{a=1}^{k}\sum_{b=a+1}^{k}\Big(H_{(a,b)}(B^{(a,b)}) + H_{(b,a)}(B^{(b,a)})\Big),$$

where $H_{(a,b)}\colon \mathbb{R}^{n_a\times n_b} \to \mathbb{R}^{N\times N}$ produces a matrix whose (a,b)th block is the input matrix, and is otherwise zero. Then

$$\begin{aligned}P_{\Lambda^\perp} B P_{\Lambda^\perp} &= \sum_{a=1}^{k}\sum_{b=a+1}^{k} P_{\Lambda^\perp}\Big(H_{(a,b)}(B^{(a,b)}) + H_{(b,a)}(B^{(b,a)})\Big)P_{\Lambda^\perp} \\ &= \sum_{a=1}^{k}\sum_{b=a+1}^{k}\Big(H_{(a,b)}(P_{1^\perp}B^{(a,b)}P_{1^\perp}) + H_{(b,a)}(P_{1^\perp}B^{(b,a)}P_{1^\perp})\Big),\end{aligned}$$

and so the triangle inequality gives

$$\begin{aligned}\|P_{\Lambda^\perp} B P_{\Lambda^\perp}\| &\leqslant \sum_{a=1}^{k}\sum_{b=a+1}^{k}\|H_{(a,b)}(P_{1^\perp}B^{(a,b)}P_{1^\perp}) + H_{(b,a)}(P_{1^\perp}B^{(b,a)}P_{1^\perp})\| \\ &= \sum_{a=1}^{k}\sum_{b=a+1}^{k}\|P_{1^\perp}B^{(a,b)}P_{1^\perp}\|,\end{aligned}$$

where the last equality can be verified by considering the spectrum of the square:

$$\begin{aligned}&\Big(H_{(a,b)}(P_{1^\perp}B^{(a,b)}P_{1^\perp}) + H_{(b,a)}(P_{1^\perp}B^{(b,a)}P_{1^\perp})\Big)^2 \\ &\qquad = H_{(a,a)}\Big((P_{1^\perp}B^{(a,b)}P_{1^\perp})(P_{1^\perp}B^{(a,b)}P_{1^\perp})^\top\Big) \\ &\qquad\qquad + H_{(b,b)}\Big((P_{1^\perp}B^{(a,b)}P_{1^\perp})^\top(P_{1^\perp}B^{(a,b)}P_{1^\perp})\Big).\end{aligned}$$

At this point, we use the definition of B (4.23) to get

$$\|P_{1^\perp} B^{(a,b)} P_{1^\perp}\| = \frac{\|P_{1^\perp} u_{(a,b)}\|_2 \|P_{1^\perp} u_{(b,a)}\|_2}{\rho_{(a,b)}}.$$

Recalling the definition of $u_{(a,b)}$ (4.23) and combining these estimates then produces the result.

4.1.3.2 Proof Theorem 4.1.9

In this section, we apply the certificate from Corollary 4.1.8 to the $(\mathcal{D}, \gamma, n)$-stochastic ball model (see Definition 1.4.1) to prove our main result. We will prove Theorem 4.1.9 with the help of several lemmas.

Lemma 4.1.10. *Denote*

$$c_a := \frac{1}{n}\sum_{i=1}^{n} x_{a,i}, \qquad \Delta_{ab} := \|\gamma_a - \gamma_b\|_2, \qquad O_{ab} := \frac{\gamma_a + \gamma_b}{2}.$$

Then the $(\mathcal{D}, \gamma, n)$-stochastic ball model satisfies the following estimates:

$$\|c_a - \gamma_a\|_2 < \epsilon \quad w.p. \ -e^{-\Omega_{m,\epsilon}(n)} \tag{4.27}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} \|r_{a,i}\|_2^2 - \mathbb{E}\|r\|_2^2\right| < \epsilon \quad w.p. \ -e^{-\Omega_{\epsilon}(n)} \tag{4.28}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} \|x_{a,i} - O_{ab}\|_2^2 - \mathbb{E}\|r + \gamma_a - O_{ab}\|_2^2\right| < \epsilon \quad w.p. \ 1 - e^{-\Omega_{\Delta_{ab},\epsilon}(n)} \tag{4.29}$$

Proof. Since $\mathbb{E}r = 0$ and $\|r\|_2^2 \leqslant 1$ almost surely, one may lift

$$X_{a,i} := \begin{bmatrix} 0 & r_{a,i}^\top \\ r_{a,i} & 0 \end{bmatrix}$$

and apply the Matrix Hoeffding inequality [61] to conclude that

$$\Pr\left(\left\|\sum_{i=1}^{n} r_{a,i}\right\|_2 \geqslant t\right) \leqslant me^{-t^2/8n}.$$

Taking $t := \epsilon n$ then gives (4.27). For (4.28) and (4.29), notice that the random variables in each sum are iid and confined to an interval almost surely, and so the result follows from Hoeffding's inequality. □

Lemma 4.1.11. *Under the* $(\mathcal{D}, \gamma, n)$*-stochastic ball model, we have* $D^{(a,b)}1 - D^{(a,a)}1 = 4np + q$*, where*

$$\begin{aligned} p_i &:= r_{a,i}^\top(\gamma_a - O_{ab}) + \frac{\Delta_{ab}^2}{4} \\ q_i &:= 2n(x_{a,i} - O_{ab})^\top\left((c_a - c_b) - (\gamma_a - \gamma_b)\right) \\ &\qquad + \left(\sum_{j=1}^{n}\|x_{b,j} - O_{ab}\|_2^2 - \sum_{j=1}^{n}\|x_{a,j} - O_{ab}\|_2^2\right) \end{aligned}$$

and $|q_i| \leqslant (6 + 2\Delta_{ab})n\epsilon$ *with probability* $1 - e^{-\Omega_{m,\Delta_{ab},\epsilon}(n)}$.

Proof. Add and subtract O_{ab} and then expand the squares to get

$$\begin{aligned} e_i^\top(D^{(a,b)}1 - D^{(a,a)}1) &= \sum_{j=1}^{n}\|x_{a,i} - x_{b,j}\|_2^2 - \sum_{j=1}^{n}\|x_{a,i} - x_{a,j}\|_2^2 \\ &= n\left(-2(x_{a,i} - O_{ab})^\top(c_b - O_{ab}) + \frac{1}{n}\sum_{j=1}^{n}\|x_{b,j} - O_{ab}\|_2^2\right) \\ &\qquad - n\left(-2(x_{a,i} - O_{ab})^\top(c_a - O_{ab}) + \frac{1}{n}\sum_{j=1}^{n}\|x_{a,j} - O_{ab}\|_2^2\right) \\ &= 2n(x_{a,i} - O_{ab})^\top(c_a - c_b) + \left(\sum_{j=1}^{n}\|x_{b,j} - O_{ab}\|_2^2 - \sum_{j=1}^{n}\|x_{a,j} - O_{ab}\|_2^2\right). \end{aligned}$$

Add and subtract $\gamma_a - \gamma_b$ to $c_a - c_b$ and distribute over the resulting sum to obtain

$$\begin{aligned} e_i^\top (D^{(a,b)} 1 - D^{(a,a)} 1) &= 2n(x_{a,i} - O_{ab})^\top (\gamma_a - \gamma_b) + q \\ &= 4n\Big(r_{a,i} + (\gamma_a - O_{ab})\Big)^\top (\gamma_a - O_{ab}) + q. \end{aligned}$$

Distributing and identifying $\|\gamma_a - O_{ab}\|_2^2 = \Delta_{ab}^2/4$ explains the definition of p. To show $|q_i| \leqslant (6 + 2\Delta_{ab})n\epsilon$, apply triangle and Cauchy–Schwarz to obtain

$$\begin{aligned} |q_i| &\leqslant \left| 2n(x_{a,i} - O_{ab})^\top \Big((c_a - c_b) - (\gamma_a - \gamma_b) \Big) \right| \\ &\qquad + \left| \sum_{j=1}^n \|x_{b,j} - O_{ab}\|_2^2 - \sum_{j=1}^n \|x_{a,j} - O_{ab}\|_2^2 \right| \\ &\leqslant 2n\Big(\|r_{a,i}\|_2 + \|\gamma_a - O_{a,b}\|_2 \Big) \Big(\|c_a - \gamma_a\|_2 + \|c_b - \gamma_b\|_2 \Big) \\ &\qquad + \left| \sum_{j=1}^n \|x_{b,j} - O_{ab}\|_2^2 - \sum_{j=1}^n \|x_{a,j} - O_{ab}\|_2^2 \right| \\ &\leqslant 2n\Big(1 + \frac{\Delta_{ab}}{2} \Big) \Big(\|c_a - \gamma_a\|_2 + \|c_b - \gamma_b\|_2 \Big) \\ &\qquad + \left| \sum_{j=1}^n \|x_{b,j} - O_{ab}\|_2^2 - \sum_{j=1}^n \|x_{a,j} - O_{ab}\|_2^2 \right|. \end{aligned}$$

To finish the argument, apply (4.27) to the first term while adding and subtracting

$$\mathbb{E}\|r + \gamma_a - O_{ab}\|_2^2 = \mathbb{E}\|r + \gamma_b - O_{ab}\|_2^2,$$

from the second and apply (4.29). □

Lemma 4.1.12. *Under the $(\mathcal{D}, \gamma, n)$-stochastic ball model, we have*

$$\left| \frac{1}{n} 1^\top D^{(a,a)} 1 - 2n\mathbb{E}\|r\|_2^2 \right| \leqslant 4n\epsilon \qquad \textit{w.p.} \qquad 1 - e^{-\Omega_{\Delta_{ab},\epsilon}(n)}.$$

Proof. Add and subtract γ_a and expand the square to get

$$\frac{1}{n}e_i^\top D^{(a,a)}1 = \frac{1}{n}\sum_{j=1}^{n}\|x_{a,i}-x_{a,j}\|_2^2 = \|r_{a,i}\|_2^2 - 2r_{a,i}^\top(c_a-\gamma_a) + \frac{1}{n}\sum_{j=1}^{n}\|r_{a,j}\|_2^2.$$

The triangle and Cauchy–Schwarz inequalities then give

$$\begin{aligned}
&\left|\frac{1}{n}1^\top D^{(a,a)}1 - 2n\mathbb{E}\|r\|_2^2\right| \\
&= \left|\sum_{i=1}^{n}\left(\|r_{a,i}\|_2^2 - 2r_{a,i}^\top(c_a-\gamma_a) + \frac{1}{n}\sum_{j=1}^{n}\|r_{a,j}\|_2^2\right) - 2n\mathbb{E}\|r\|_2^2\right| \\
&\leqslant n\left|\frac{1}{n}\sum_{i=1}^{n}\|r_{a,i}\|_2^2 - \mathbb{E}\|r\|_2^2\right| + 2\sum_{i=1}^{n}|r_{a,i}^\top(c_a-\gamma_a)| + n\left|\frac{1}{n}\sum_{j=1}^{n}\|r_{a,j}\|_2^2 - \mathbb{E}\|r\|_2^2\right| \\
&\leqslant n\left|\frac{1}{n}\sum_{i=1}^{n}\|r_{a,i}\|_2^2 - \mathbb{E}\|r\|_2^2\right| + 2\sum_{i=1}^{n}\|c_a-\gamma_a\|_2 + n\left|\frac{1}{n}\sum_{j=1}^{n}\|r_{a,j}\|_2^2 - \mathbb{E}\|r\|_2^2\right| \\
&\leqslant 4n\epsilon,
\end{aligned}$$

where the last step occurs with probability $1-e^{-\Omega_{\Delta_{ab},\epsilon}(n)}$ by a union bound over (4.28) and (4.27). □

Lemma 4.1.13. *Under the $(\mathcal{D},\gamma,n)$-stochastic ball model, we have*

$$1^\top D^{(a,b)}1 - 1^\top D^{(a,a)}1 \geqslant n^2\Delta_{ab}^2 - (6+4\Delta_{ab})n^2\epsilon \quad \textit{w.p.} \quad 1-e^{-\Omega_{m,\Delta_{ab},\epsilon}(n)}.$$

Proof. Lemma 4.1.11 gives

$$\begin{aligned}
1^\top D^{(a,b)}1 - 1^\top D^{(a,a)}1 &= 1^\top(4np+q) \\
&\geqslant 4n\sum_{i=1}^{n}\left(r_{a,i}^\top(\gamma_a-O_{ab}) + \frac{\Delta_{ab}^2}{4}\right) - (6+2\Delta_{ab})n^2\epsilon \\
&\geqslant 4n\left(n(c_a-\gamma_a)^\top(\gamma_a-O_{ab}) + \frac{n\Delta_{ab}^2}{4}\right) - (6+2\Delta_{ab})n^2\epsilon.
\end{aligned}$$

Cauchy–Schwarz along with (4.27) then gives the result. □

Lemma 4.1.14. *Under the $(\mathcal{D},\gamma,n)$-stochastic ball model, there exists $C = C(\gamma)$ such that*

$$\min_{\substack{a,b\in\{1,\dots,k\}\\ a\neq b}} \min(M^{(a,b)}1) \geqslant n\Delta(\Delta-2) + Cn\epsilon \qquad \textit{w.p.} \qquad 1-e^{-\Omega_{m,\gamma,\epsilon}(n)},$$

where $\Delta := \min_{\substack{a,b\in\{1,\dots,k\}\\ a\neq b}} \Delta_{ab}$.

Proof. Fix a and b. Then by Lemma 4.1.11, the following holds with probability $1-e^{-\Omega_{m,\Delta_{ab},\epsilon}(n)}$:

$$\begin{aligned}\min\left(D^{(a,b)}1 - D^{(a,a)}1\right) &\geqslant 4n \min_{i\in\{1,\dots,n\}} \left(r_{a,i}^\top(\gamma_a - O_{ab}) + \frac{\Delta_{ab}^2}{4}\right) \\ &\qquad - (6+2\Delta_{ab})n\epsilon \\ &\geqslant n\Delta_{ab}^2 - 2n\Delta_{ab} - (6+2\Delta_{ab})n\epsilon,\end{aligned}$$

where the last step is by Cauchy–Schwarz. Taking a union bound with Lemma 4.1.12 then gives

$$\begin{aligned}&\min(M^{(a,b)}1) \\ &\quad= \min\left(D^{(a,b)}1 - D^{(a,a)}1\right) + \frac{1}{2}\left(\frac{1}{n}1^\top D^{(a,a)}1 - \frac{1}{n}1^\top D^{(b,b)}1\right) \\ &\quad\geqslant \min\left(D^{(a,b)}1 - D^{(a,a)}1\right) \\ &\qquad\qquad - \frac{1}{2}\left(\left|\frac{1}{n}1^\top D^{(a,a)}1 - 2n\mathbb{E}\|r\|_2^2\right| + \left|\frac{1}{n}1^\top D^{(b,b)}1 - 2n\mathbb{E}\|r\|_2^2\right|\right) \\ &\quad\geqslant n\Delta_{ab}(\Delta_{ab}-2) - (10+2\Delta_{ab})n\epsilon\end{aligned}$$

with probability $1-e^{-\Omega_{\Delta_{ab},\epsilon}(n)}$. The result then follows from a union bound over a and b. □

Lemma 4.1.15. *Suppose $\epsilon \leqslant 1$. Then there exists $C = C(\Delta_{ab}, m)$ such that under the $(\mathcal{D}, \gamma, n)$-stochastic ball model, we have*

$$\|P_{1^\perp} M^{(a,b)} 1\|_2^2 \leqslant \frac{4n^3 \Delta_{ab}^2}{m} + Cn^3\epsilon$$

with probability $1 - e^{-\Omega_{m,\Delta_{ab},\epsilon}(n)}$.

Proof. First, a quick calculation reveals

$$\begin{aligned}
e_i^\top M^{(a,b)} 1 &= e_i^\top D^{(a,b)} 1 - e_i^\top D^{(a,a)} 1 + \frac{1}{2}\left(\frac{1}{n} 1^\top D^{(a,a)} 1 - \frac{1}{n} 1^\top D^{(b,b)} 1\right), \\
\frac{1}{n} 1^\top M^{(a,b)} 1 &= \frac{1}{n} 1^\top D^{(a,b)} 1 - \frac{1}{2}\left(\frac{1}{n} 1^\top D^{(a,a)} 1 + \frac{1}{n} 1^\top D^{(b,b)} 1\right),
\end{aligned}$$

from which it follows that

$$\begin{aligned}
e_i^\top P_{1^\perp} M^{(a,b)} 1 &= e_i^\top M^{(a,b)} 1 - \frac{1}{n} 1^\top M^{(a,b)} 1 \\
&= \left(e_i^\top D^{(a,b)} 1 - \frac{1}{n} 1^\top D^{(a,b)} 1\right) - \left(e_i^\top D^{(a,a)} 1 - \frac{1}{n} 1^\top D^{(a,a)} 1\right) \\
&= e_i^\top P_{1^\perp} (D^{(a,b)} 1 - D^{(a,a)} 1).
\end{aligned}$$

As such, we have

$$\begin{aligned}
\|P_{1^\perp} M^{(a,b)} 1\|_2^2 &= \|P_{1^\perp} (D^{(a,b)} 1 - D^{(a,a)} 1)\|_2^2 \\
&= \|D^{(a,b)} 1 - D^{(a,a)} 1\|_2^2 - \|P_1 (D^{(a,b)} 1 - D^{(a,a)} 1)\|_2^2. \qquad (4.30)
\end{aligned}$$

To bound the first term, we apply the triangle inequality over Lemma 4.1.11:

$$\|D^{(a,b)} 1 - D^{(a,a)} 1\|_2 \leqslant 4n\|p\|_2 + \|q\|_2 \leqslant 4n\|p\|_2 + (6 + 2\Delta_{ab}) n^{3/2} \epsilon. \quad (4.31)$$

We proceed by bounding $\|p\|_2$. To this end, note that the p_i's are iid random variables whose outcomes lie in a finite interval (of width determined by Δ_{ab}) with probability 1. As such, Hoeffding's inequality gives

$$\left|\frac{1}{n}\sum_{i=1}^{n} p_i^2 - \mathbb{E}p_1^2\right| \leqslant \epsilon \qquad \text{w.p.} \qquad 1 - e^{-\Omega_{\Delta_{ab},\epsilon}(n)}.$$

With this, we then have

$$\|p\|_2^2 = n\left(\frac{1}{n}\sum_{i=1}^{n} p_i^2 - \mathbb{E}p_1^2 + \mathbb{E}p_1^2\right) \leqslant n\mathbb{E}p_1^2 + n\epsilon \tag{4.32}$$

in the same event. To determine $\mathbb{E}p_1^2$, first take $r_1 := e_1^\top r$. Then since the distribution of r is rotation invariant, we may write

$$p_1 = r_{a,1}^\top(\gamma_a - O_{ab}) + \|\gamma_a - O_{ab}\|_2^2 = \frac{\Delta_{ab}}{2}r_1 + \frac{\Delta_{ab}^2}{4},$$

where the second equality above is equality in distribution. We then have

$$\mathbb{E}p_1^2 = \mathbb{E}\left(\frac{\Delta_{ab}}{2}r_1 + \frac{\Delta_{ab}^2}{4}\right)^2 = \frac{\Delta_{ab}^2}{4}\mathbb{E}r_1^2 + \frac{\Delta_{ab}^4}{16}. \tag{4.33}$$

We also note that $1 \geqslant \mathbb{E}\|r\|_2^2 = m\mathbb{E}r_1^2$ by linearity of expectation, and so

$$\mathbb{E}r_1^2 \leqslant \frac{1}{m}. \tag{4.34}$$

Combining (4.31), (4.32), (4.33) and (4.34) then gives

$$\begin{aligned}\|D^{(a,b)}1 - D^{(a,a)}1\|_2 \\ \leqslant \left(\frac{4n^3\Delta_{ab}^2}{m} + n^3\Delta_{ab}^4 + 16n^3\epsilon\right)^{1/2} + (6 + 2\Delta_{ab})n^{3/2}\epsilon.\end{aligned} \tag{4.35}$$

To bound the second term of (4.30), first note that

$$\|P_1(D^{(a,b)}1 - D^{(a,a)}1)\|_2 = \frac{1}{\sqrt{n}}\left|1^\top D^{(a,b)}1 - 1^\top D^{(a,a)}1\right|. \tag{4.36}$$

Lemma 4.1.13 then gives

$$\begin{aligned}\left|1^\top D^{(a,b)}1 - 1^\top D^{(a,a)}1\right| &\geqslant 1^\top D^{(a,b)}1 - 1^\top D^{(a,a)}1 \\ &\geqslant n^2\Delta_{ab}^2 - (6 + 4\Delta_{ab})n^2\epsilon \end{aligned} \tag{4.37}$$

with probability $1 - e^{-\Omega_{m,\Delta_{ab},\epsilon}(n)}$. Using (4.30) to combine (4.35) with (4.36) and (4.37) then gives the result. □

Lemma 4.1.16. *There exists $C = C(\gamma)$ such that under the $(\mathcal{D},\gamma,n)$-stochastic ball model, we have*

$$\rho_{(a,b)} \geqslant n^2\big(\Delta_{ab}^2 - \Delta(\Delta-2)\big) - Cn^2\epsilon \qquad \textit{w.p.} \qquad 1 - e^{-\Omega_{\mathcal{D},\gamma,\epsilon}(n)}.$$

Proof. Recall from (4.23) that

$$\begin{aligned}\rho_{(a,b)} = u_{(a,b)}^\top 1 &= 1^\top M^{(a,b)}1 - nz \\ &= 1^\top M^{(a,b)}1 - n\min_{\substack{a,b\in\{1,\dots,k\}\\ a\neq b}} \min(M^{(a,b)}1).\end{aligned} \tag{4.38}$$

To bound the first term, we leverage Lemma 4.1.13:

$$\begin{aligned}1^\top M^{(a,b)}1 &= 1^\top D^{(a,b)}1 - \frac{1}{2}(1^\top D^{(a,a)}1 + 1^\top D^{(b,b)}1) \\ &= \frac{1}{2}\Big(1^\top D^{(a,b)}1 - 1^\top D^{(a,a)}1\Big) + \frac{1}{2}\Big(1^\top D^{(b,a)}1 - 1^\top D^{(b,b)}1\Big) \\ &\geqslant n^2\Delta_{ab}^2 - (6 + 4\Delta_{ab})n^2\epsilon\end{aligned}$$

with probability $1 - e^{-\Omega_{m,\Delta_{ab},\epsilon}(n)}$. To bound the second term in (4.38), note from Lemma 4.1.12 that

$$\begin{aligned}
&\min(M^{(a,b)}1) \\
&\quad = \min\left(D^{(a,b)}1 - D^{(a,a)}1\right) + \frac{1}{2}\left(\frac{1}{n}1^\top D^{(a,a)}1 - \frac{1}{n}1^\top D^{(b,b)}1\right) \\
&\quad \leqslant \min\left(D^{(a,b)}1 - D^{(a,a)}1\right) \\
&\qquad\qquad + \frac{1}{2}\left(\left|\frac{1}{n}1^\top D^{(a,a)}1 - 2n\mathbb{E}\|r\|_2^2\right| + \left|\frac{1}{n}1^\top D^{(b,b)}1 - 2n\mathbb{E}\|r\|_2^2\right|\right) \\
&\quad \leqslant \min\left(D^{(a,b)}1 - D^{(a,a)}1\right) + 4n\epsilon
\end{aligned}$$

with probability $1 - e^{-\Omega_{\Delta_{ab},\epsilon}(n)}$. Next, Lemma 4.1.11 gives

$$\min\left(D^{(a,b)}1 - D^{(a,a)}1\right) \leqslant n\Delta_{ab}^2 + (6 + 2\Delta_{ab})n\epsilon + 4n \min_{i\in\{1,\dots,n\}} r_{a,i}^\top(\gamma_a - O_{ab}).$$

By assumption, we know $\|r\|_2 \geqslant 1 - \epsilon$ with positive probability regardless of $\epsilon > 0$. It then follows that

$$r^\top(\gamma_a - O_{ab}) \leqslant -\frac{\Delta_{ab}}{2} + \epsilon$$

with some (ϵ-dependent) positive probability. As such, we may conclude that

$$\min_{i\in\{1,\dots,n\}} r_{a,i}^\top(\gamma_a - O_{ab}) \leqslant -\frac{\Delta_{ab}}{2} + \epsilon \qquad \text{w.p.} \qquad 1 - e^{-\Omega_{\mathcal{D},\epsilon}(n)}.$$

Combining these estimates then gives

$$\min(M^{(a,b)}1) \leqslant n\Delta_{ab}^2 - 2n\Delta_{ab} + (10 + 2\Delta_{ab})n\epsilon \quad \text{w.p.} \quad 1 - e^{-\Omega_{\mathcal{D},\Delta_{ab},\epsilon}(n)}.$$

Performing a union bound over a and b then gives

$$\min_{\substack{a,b\in\{1,\dots,k\}\\ a\neq b}} \min(M^{(a,b)}1) \leqslant n\Delta^2 - 2n\Delta + (10 + 2\Delta)n\epsilon \text{ w.p. } 1 - e^{-\Omega_{\mathcal{D},\gamma,\epsilon}(n)}.$$

Combining these estimates then gives the result. □

Lemma 4.1.17. *Under the $(\mathcal{D},\gamma,n)$-stochastic ball model, we have*

$$\|\Psi\| \leqslant \left(\frac{(1+\epsilon)\sigma}{\sqrt{m}} + \epsilon\right)\sqrt{N} \qquad \textit{w.p.} \qquad 1 - e^{-\Omega_{m,k,\sigma,\epsilon}(n)},$$

where $\sigma^2 := \mathbb{E}\|r\|_2^2$ for $r \sim \mathcal{D}$.

Proof. Let R denote the matrix whose (a,i)th column is $r_{a,i}$. Then

$$\Psi = R - \left[(c_1-\gamma_1)1^\top \ \cdots \ (c_k-\gamma_k)1^\top\right],$$

and so the triangle inequality gives

$$\|\Psi\| \leqslant \|R\| + \left\|\left[(c_1-\gamma_1)1^\top \ \cdots \ (c_k-\gamma_k)1^\top\right]\right\| \leqslant \|R\| + \left(n\sum_{a=1}^{k}\|c_a-\gamma_a\|_2^2\right)^{1/2}$$

where the last estimate passes to the Frobenius norm. For the first term, since $\mathcal{D}$ is rotation invariant, we may apply Theorem 5.41 in [64]:

$$\|R\| \leqslant (1+\epsilon)\sigma\sqrt{\frac{N}{m}} \qquad \text{w.p.} \qquad 1 - e^{-\Omega_{m,\sigma,\epsilon}(n)}.$$

For the second term, apply (4.27). The union bound then gives the result. □

Proof of Theorem 4.1.9. First, we combine Lemmas 4.1.15, 4.1.16 and 4.1.17: For every $\delta > 0$, there exists an $\epsilon > 0$ such that

$$\begin{aligned}
2\|\Psi\|^2 + \sum_{a=1}^{k}\sum_{b=a+1}^{k}\frac{\|P_{1^\perp}M^{(a,b)}1\|_2\|P_{1^\perp}M^{(b,a)}1\|_2}{\rho_{(a,b)}} & \\
\leqslant 2\left(\frac{1+\epsilon}{\sqrt{m}}+\epsilon\right)^2 nk + \sum_{a=1}^{k}\sum_{b=a+1}^{k}\frac{4n^3\Delta_{ab}^2/m + Cn^3\epsilon}{n^2(\Delta_{ab}^2 - \Delta(\Delta-2)) - Cn^2\epsilon} & \\
\leqslant n\left(\frac{2k}{m} + \frac{4}{m}\sum_{a=1}^{k}\sum_{b=a+1}^{k}\frac{\Delta_{ab}^2}{\Delta_{ab}^2 - \Delta(\Delta-2)} + \delta\right) & \qquad (4.39)
\end{aligned}$$

with probability $1 - e^{-\Omega_{\mathcal{D},\gamma,\epsilon}(n)}$. Next, the uniform bound $\Delta_{ab} \geqslant \Delta$ implies

$$\frac{\Delta_{ab}^2}{\Delta_{ab}^2 - \Delta(\Delta-2)} = \frac{1}{1 - \Delta(\Delta-2)/\Delta_{ab}^2} \leqslant \frac{1}{1 - \Delta(\Delta-2)/\Delta^2} = \frac{\Delta}{2}.$$

Combining this with (4.39) and considering Lemma 4.1.14, it then suffices to have

$$\frac{2k}{m} + \frac{4}{m} \cdot \binom{k}{2} \cdot \frac{\Delta}{2} < \Delta(\Delta-2).$$

Rearranging then gives

$$\Delta > 2 + \frac{2k}{m\Delta} + \frac{k(k-1)}{m},$$

which is implied by the hypothesis since $\Delta \geqslant 2$. □

4.2 Integrality for the k-medians LP relaxation

Recall the k-medians linear programming formulation (k-medians lp). Using the techniques explained in Chapter 2 we compute its dual linear program, which is given in (k-medians lp dual). In this section we construct a dual certificate and we use it to prove that the k-medians LP recovers the planted clusters if they satisfy certain deterministic conditions (that we call separation and center dominance). We then prove that the deterministic conditions are satisfied for the stochastic ball model under minimal separation with high probability.

In this section we use the following notation: $P = \{x_1, \ldots, x_N\}$ is a subset of a metric space (X, d). Letters p, q, s will denote points in P. $A_1, \ldots, A_k$

will denote the planted clusters with respective k-medians centers $c_1, \ldots, c_k$.
With this notation we have the linear programs (k-medians lp) and (k-medians lp dual).

$$\begin{aligned} \underset{z \in \mathbb{R}^{N \times N}, y \in \mathbb{R}^N}{\text{minimize}} \quad & \sum_{p \in P} \sum_{q \in P} d(p,q) z_{pq} && \text{(k-medians lp)} \\ \text{subject to} \quad & \sum_{p=1}^{n} z_{pq} = 1 \ \forall q \in P, \ z_{pq} \leqslant y_p \ \forall p, q \in P, \\ & \sum_{p \in P} y_p = k, \ z_{pq}, y_p \in [0,1]. \end{aligned}$$

$$\begin{aligned} \underset{\xi \in \mathbb{R}, \alpha \in \mathbb{R}^N, \beta \in \mathbb{R}^{N \times N}}{\text{maximize}} \quad & \sum_{q \in P} \alpha_q - k\xi && \text{(k-medians lp dual)} \\ \text{subject to} \quad & \alpha_q \leqslant \beta_{pq} + d(p,q) \ \ \forall p, q \in P \\ & \sum_{q \in P} \beta_{pq} \leqslant \xi, \ \forall p \in P, \quad \beta_{pq} \geqslant 0 \ \ \forall p, q \in P \end{aligned}$$

The solution $z \in \mathbb{R}^{N \times N}$ of (k-medians lp) is a clustering if and only if it is integral (i.e. z_{pq} are integers for all $p, q \in P$). In this setting the variable $y_p \in \{0,1\}$ indicates whether the point $p \in P$ is a center or not. The variable $z_{pq} \in \{0,1\}$ for $p, q \in P$ indicates whether or not the point p is the center for the point q. Each point has a unique center, and a cluster is the set of points sharing the same center.

Note that the solution of (k-medians lp) and (k-medians) are generically unique since no constraint is parallel to the objective function, hence motivating the following definitions.

Definition 4.2.1. For $A_t \subseteq P$, let c_t the center of A_t

$$c_t = \operatorname*{argmin}_{p \in A_t} \sum_{q \in A_t} d(p,q), \text{ and } \mathrm{OPT}_t = \min_{p \in A_t} \sum_{q \in A_t} d(p,q).$$

We prove optimality of a particular integral solution to (k-medians lp) by showing there exists a dual feasible solution to (k-medians lp dual) whose dual objective value matches the primal objective value of the intended integral solution - a so-called *dual certificate*. When the solution of (k-medians lp) is integral, it is also degenerate, since most of the variables are zero. In fact we experimentally observed that the dual has multiple solutions. Indeed, motivated by this observation and experimental evidence, we can essentially enforce an extra constraint in the dual by asking that the variables α be constant within each cluster. Given α's as such, the β's and ξ's are then easily identified. We now formulate a sufficient condition for integrality based on these observations:

Lemma 4.2.1. *Consider sets $A_1, \ldots, A_k$ with $n_1, \ldots, n_k$ points respectively. If $\exists \alpha_1, \ldots, \alpha_k$ s.t for each $s \in A_1 \cup \ldots \cup A_k$,*

$$\frac{1}{k}\left(\sum_{t=1}^{k}\left[n_t\alpha_t - \min_{p \in A_t}\sum_{q \in A_t} d(p,q)\right]\right) \geqslant \sum_{q \in A_1}(\alpha_1 - d(s,q))_+ + \ldots + \sum_{q \in A_k}(\alpha_k - d(s,q))_+, \quad (4.40)$$

then the k-medians LP (k-medians lp) is integral and the partition in clusters $A_1, \ldots, A_k$ is optimal.

Proof. By strong duality, the intended cluster solution is optimal if the corresponding LP objective value

$$\min_{p \in A_1} \sum_{q \in A_1} d(p,q) + \ldots + \min_{p \in A_k} \sum_{q \in A_k} d(p,q)$$

is less than or equal to the dual objective for some feasible dual variables. By restricting the dual variables α_q to be constant within each cluster, and by setting ξ to be equal to the RHS of the Lemma statement, a computation verifies that the dual objective is at least the cost of the intended clustering. Moreover, it is also easy to see that for this setting of ξ and α_q's, the dual constraints are trivially satisfied. □

Note that the sufficient condition in Lemma 4.2.1 is similar to the sufficient condition considered in [50], but turns out to be more useful since it allows us to get down to optimal cluster separation $\Delta = 2 + \epsilon$ (whereas in [50] the authors prove the result for $\Delta \geqslant 3.75$).

A possible interpretation for the dual variables (which has been exploited by the primal-dual based approximation algorithms for the k-medians problem in [8]) is as distance thresholds. In the RHS of equation (4.40) in $\sum_{q \in A_t} (\alpha_t - d(s,q))_+$ a point $s \in P$ gets positive contribution from points $q \in A_t$ that are at a distance smaller than α_t. In this sense, a point in the set A_t can only "see" other points within a distance α_t.

Following this intuition, one way to prove that inequality (4.40) holds is to show that we can choose feasible dual variables $\alpha_1, \ldots, \alpha_k$ to satisfy

- Each center sees exactly its own cluster i.e. $(\alpha_t - d(c_t, q))_+ > 0$ if and only if $q \in A_t$.
- The RHS of (4.40) attains its maximum in the centers $c_1, \ldots, c_k$.
- Each of the terms $n_t \alpha_t - \min_{p \in A_t} \sum_{q \in A_t} d(p, q)$ in the average in the LHS of (4.40) are the same.

Our strategy is to provide a set of conditions in our data points that guarantee such feasible dual variables exist. Assume the sets $A_1, \ldots, A_k$ are contained in disjoint balls $B_{r_1}(c_1), \ldots, B_{r_k}(c_k)$ respectively (where we use the notation $B_r(c)$ to indicate a ball of radius r centered at c). Here we do a slight abuse of notation, the points $c_1, \ldots, c_k$ do not need to be the k-medians centers (they could be the k-medians centers, centroids, or any point in the cluster that satisfies the conditions). However we will prove that the k-medians centers can be chosen to be such $c_1, \ldots, c_k$.

Suppose that $\alpha_1, \ldots, \alpha_k$, $\alpha_t > r_t$, are such that for all $a \neq b$, $B_{\alpha_a}(c_a) \cap B_{\alpha_b}(c_b) = \emptyset$. Given the α's there exist $\tau_1, \ldots, \tau_k > 0$ sufficiently small that any $x \in B_{\tau_t}(c_t)$ is seen only by points in its own ball (see Definition 4.2.3 for a precise statement). We now define conditions on the sets $A_1, \ldots, A_k$ which imply integrality of the linear programming relaxation (k-medians lp). Note the conditions can be expressed for generic radius and number of points, but for simplicity, we assume for the remainder of the section $n_1 = \ldots = n_k = n$ and $r_1 = \ldots = r_k = 1$ (as in the stochastic ball model). Roughly speaking, our conditions ask that (a) The clusters are separated, being contained in

disjoint balls, (b) Outside of a certain neighborhood of the center, no point is a good center for its own cluster and (c) No point gets too much contribution from any other cluster. More precisely, we require the following *separation* and *center dominance* conditions:

Definition 4.2.2 (Separation)**.** Let the sets $A_1, \ldots, A_k$ in X, such that

$$\mathrm{OPT}_1 \leqslant \ldots \leqslant \mathrm{OPT}_k$$

We say such sets satisfy the separation condition if they are included in k disjoint balls: $A_1 \subset B_1(c_1)$, $\ldots$, $A_k \subset B_1(c_k)$, $d(c_a, c_b) = 2 + \delta_{ab}$ for $a \neq b$ where $\delta_{ab} > 0$, and the distance between c_a and c_b satisfies:

$$\Theta := \min_{1 \leqslant a,b \leqslant k} \delta_{ab} > \frac{\mathrm{OPT}_k - \mathrm{OPT}_1}{n} \tag{4.41}$$

Remark 4.2.1. The expression $\frac{\mathrm{OPT}_k - \mathrm{OPT}_1}{n}$ provides a way of measuring how different the clusters are from each other. For example, if the clusters are symmetric, then $\frac{\mathrm{OPT}_k - \mathrm{OPT}_1}{n} = 0$. This condition requires bigger separation when clusters are different.

We also require a *center dominance* condition. Consider the contribution function $P^{(\alpha_1,\ldots,\alpha_k)} : X \to \mathbb{R}$ as the sum of all contributions that a point can get:

$$P^{(\alpha_1,\ldots,\alpha_k)}(y) = \sum_{i=1}^{k} \sum_{x \in A_t} (\alpha_t - d(y,x))_+.$$

The center dominance condition essentially says that the contribution function attains its maximum in a small neighborhood of the center of each ball, as long as the parameters α are chosen from some small interval.

Definition 4.2.3 (Center dominance). $A_1, \ldots, A_k$ satisfy center dominance in the interval $(a, b) \subset (1, 1 + \Theta)$ if

$$b - a > \frac{OPT_k - OPT_1}{n} \tag{4.42}$$

and for all $\alpha_1, \ldots, \alpha_k \in (a, b)$ there exist $\tau_1, \ldots, \tau_k > 0$ such that for all $x \in B_{\tau_t}(c_t)$, $t = 1, \ldots, k$

$$B_{\alpha_a}(x) \cap B_{r_a}(c_a) = \begin{cases} B_{r_t}(c_t) & \text{if } a = t \\ \emptyset & \text{otherwise} \end{cases} \tag{4.43}$$

$$\max_{y \in A_t \setminus B_{\tau_t}(c_t)} P^{(\alpha_1, \ldots, \alpha_k)}(y) < \max_{y \in B_{\tau_t}(c_t)} P^{(\alpha_1, \ldots, \alpha_k)}(y) \tag{4.44}$$

Theorem 4.2.2 states deterministic conditions that if satisfied by P, (k-medians lp) recovers the planted clusters. The proof of this theorem is in Section 4.2.2.

Theorem 4.2.2. *If $A_1, \ldots, A_k$ are k sets in a metric space (X, d) satisfying separation and center dominance, then there is an integral solution of* (k-medians lp) *and it corresponds to separating $P = A_1 \cup \ldots \cup A_k$ in the clusters $A_1, \ldots, A_k$.*

Indeed, a broad class of distributions are likely to satisfy these conditions. The following theorem shows that with high probability under the stochastic ball model. The proof specifically requires that the probability of any ball containing 0 is positive.

Theorem 4.2.3. *If points are drawn from a (μ, γ, n)-stochastic ball model with $\Delta > 2$ then, for all $\gamma < 1$, there exists N_0 such that, if $n > N_0$, then the solution of* (k-medians lp) *is integral with probability at least γ.*

The proof of this theorem is in Section 4.2.1. The main idea is that given k balls with the same continuous probability distribution, for large values of n, the separation condition is just a consequence of the weak law of large numbers. And one can prove that center dominance holds in expectation, so it will hold with high probability if the number of points n is large enough due concentration of measure. Note that the condition that all measures are the same and rotationally symmetric can be dropped as long as the expectation of the contribution function attains its maximum in a point close enough to the center of the ball and $\lim_{n\to\infty} \frac{OPT_k - OPT_1}{n} < d(c_a, c_b) - 2$ for all $a \neq b$.

4.2.1 Proof of Theorem 4.2.2

Proof. Recall Lemma 4.2.1. We need to show there exists $\alpha_1, \ldots, \alpha_k$ such that for each $s \in A_1 \cup \ldots \cup A_k$ equation (4.40) holds:

$$\frac{1}{k}\left(n_1\alpha_1 - \min_{p\in A_1} \sum_{q\in A_1} d(p,q) + \ldots + n_k\alpha_k - \min_{p\in A_k} \sum_{q\in A_k} d(p,q) \right) \geqslant \sum_{q\in A_1} (\alpha_1 - d(s,q))_+ + \ldots + \sum_{q\in A_k} (\alpha_k - d(s,q))_+$$

First, note that by the center dominance property (Definition 4.2.3), that among all points within a cluster A_t, the maximum RHS is attained for $s \in B_{\tau_t}(c_t)$, i.e., for s in a small ball around c_t. Moreover, from the separation property (Definition 4.2.2), it is easy to see that points in $B_{\tau_t}(c_t)$ don't receive any contribution (in the LHS) from points in other clusters, therefore

the following holds:

$$\begin{aligned}
\max_{s\in A_t}\sum_{q\in A_1}(\alpha_1 - d(s,q))_+ + \ldots + \sum_{q\in A_k}(\alpha_k - d(s,q))_+ & \\
&= \max_{s\in B_{\tau_t}(c_t)}\sum_{q\in A_t}\alpha_t - d(s,q) \\
&= n_t\alpha_t - \sum_{q\in A_t} d(s,q) \\
&\leqslant n_t\alpha_t - \min_{p\in A_t}\sum_{q\in A_t} d(p,q) \\
&= n_t\alpha_t - \mathrm{OPT}_t
\end{aligned}$$

Now, the RHS of (4.40) maximizes s over all clusters t, so we additionally enforce that every element in the average on the LHS is the same:

$$n_1\alpha_1 - \mathrm{OPT}_1 = n_2\alpha_2 - \mathrm{OPT}_2 = \ldots = n_k\alpha_k - \mathrm{OPT}_k \tag{4.45}$$

Under this condition, it is easy to see that (4.40) holds for all $s \in A_1 \cup \ldots \cup A_k$. Since the points and the sets are given, this is a system of linear equations with one degree of freedom. □

4.2.2 Proof of Theorem 4.2.3

Proof sketch. The proof of this theorem consists of showing that separation and central dominance conditions holds with high probability when the points are drawn from the stochastic ball model.

Step 0 Let $c_t = \gamma_t$ the geometric center of the ball. For $z \in \bigcup_{t=1}^{k} B_1(c_t)$ and

$(\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k$ let the random variable

$$P^{(\alpha_1,\ldots,\alpha_k)}(z) = \sum_{t=1}^{k} \sum_{x_{t,i} \in A_t} (\alpha_t - d(z, x_{t,i}))_+ = \sum_{i=1}^{n} P_i^{(\alpha_1,\ldots,\alpha_k)}(z) \text{ where}$$

$$P_i^{(\alpha_1,\ldots,\alpha_k)}(z) = \sum_{t=1}^{k} (\alpha_t - d(z, x_{t,i}))_+$$

We need to show that for some $\alpha_1, \ldots, \alpha_k$ satisfying (4.45) the maximum of $\left\{ P^{(\alpha_1,\ldots,\alpha_k)}(x_{t,i}) \right\}_{i=1}^{n}$ is attained in some $x_{t,i} \in B_{\tau_t}(c_t)$ for every $t = 1, \ldots, k$ with high probability.

Step 1 First we show that for certain $\alpha = \alpha_1 = \ldots = \alpha_k$, the function $\mathbb{E}P_i^{(\alpha,\ldots,\alpha)}(z)$ restricted to $z \in B_1(c_t)$ attains its maximum at $z = c_t$ for all $t = 1, \ldots, k$.

The proof is done in Lemma 4.2.4. This is the step where we use that the measure is rotationally symmetric. In fact, this assumption is not strictly needed: any continuous probability distribution that satisfies the thesis of Step 1 and has positive probability in every neighborhood of the center would guarantee asymptotic recovery.

Step 2 In Lemma 4.2.5 we use that $P_i^{(\alpha_1,\ldots,\alpha_k)}(z)$ is continuous with respect to $(\alpha_1, \ldots, \alpha_k)$ and μ_t is continuous with respect to the Lebesgue measure to show that there exists some $\upsilon > 0$ with the following property: if $\alpha_1, \ldots, \alpha_k \in (\alpha - \upsilon, \alpha + \upsilon)$ then the maximum of $\mathbb{E}P_i^{(\alpha_1,\ldots,\alpha_k)}(z)$ restricted to $B_1(c_t)$ is attained at $z = c_t$.

Step 3 The weak law of large numbers implies that for all $a, b \in \{1, \ldots, k\}$,

the random variable $\frac{OPT_a - OPT_b}{n}$ converges to zero in probability, i.e.:

$$\text{For every } \nu > 0, \quad \lim_{n\to\infty} \Pr\left(\left|\frac{OPT_a - OPT_b}{n}\right| < \nu\right) = 1$$

For every $\gamma_0 < 1$ if we have n large enough, we can assure that with probability greater than γ_0, $\alpha_1, \ldots, \alpha_k$ can be chosen to be in $(\alpha - \upsilon, \alpha + \upsilon)$. In particular for $(\alpha_1, \ldots, \alpha_k)$ satisfying (4.45) the maximum of $\mathbb{E}P_i^{(\alpha_1,\ldots,\alpha_k)}(z)$ restricted to $B_1(c_t)$ is attained at $z = c_t$.

Step 4 In Lemma 4.2.6 we use concentration inequalities to convert the claim in Step 3 about $\mathbb{E}P_i^{(\alpha_1,\ldots,\alpha_k)}(z)$ to the claim we need to show about $P^{(\alpha_1,\ldots,\alpha_k)}(z)$ with high probability. Given $\gamma_1 < 1$ if the number of points n is large enough, and the probability of having a point close to the center of the ball is greater than zero, then with probability greater than γ_1, the maximum of $\left\{P^{(\alpha_1,\ldots,\alpha_k)}(x_{t,i})\right\}_{i=1}^{n}$ is attained in some $x_{t,i} \in B_{\tau_t}(c_t)$ for every $t = 1, \ldots, k$. Which proves the theorem.

□

Lemma 4.2.4. *In the hypothesis of Theorem 4.2.3 there exists $\alpha > 1$ such that for all $j = 1, \ldots, k$, $\mathbb{E}P^{(\alpha,\ldots,\alpha)}(z)$ restricted to $z \in B_1(c_j)$ attains its maximum in $z = c_j$.*

Proof. Let $z \in B_1(c_j)$. Note that

$$\begin{aligned}\mathbb{E}P^{(\alpha,\ldots,\alpha)}(z) &= n\mathbb{E}P_i^{(\alpha,\ldots,\alpha)}(z)\\ &= n\left(\int_{B_1(c_j)\cap B_\alpha(z)} \alpha - d(x,z)\,d\mu_j x + \sum_{t\neq j}\int_{B_1(c_t)\cap B_\alpha(z)} \alpha - d(x,z)\,d\mu_t x\right).\end{aligned}$$

Define $\alpha(z) > 1$ the maximum value of alpha such that $B_\alpha(z) \cap \bigcup_{t \neq j} B_1(c_t)$ can be copied isometrically inside $B_1(c_j)$ along the boundary without intersecting each other and without intersecting $B_\alpha(z)$ as demonstrated in Figure 4.1. Let $\alpha = \max\{\alpha(z) : z \in \cup_{j=1}^{k} B_1(c_j)\}$. We know $\alpha > 1$ since the balls are separated: $d(c_t, c_j) > 2$ whenever $t \neq j$.

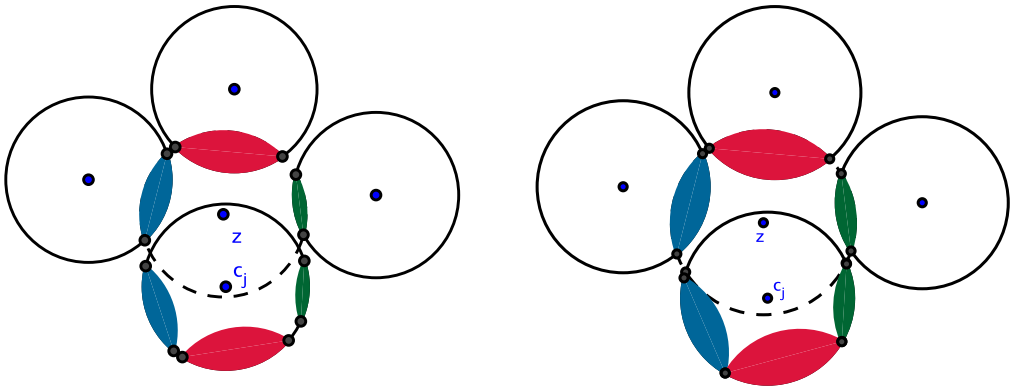


Figure 4.1: Illustration for proof of Lemma 4.2.4.
Let the circles $B_1(c_t)$ be represented by the solid lined circles and the dashed lined circle be $B_\alpha(z)$. In the left image, $\alpha = 1$. Since the circles $B_1(c_t)$ do not intersect each other, then we can consider $B_\alpha(z) \cap \bigcup_{t \neq j} B_1(c_t)$ copied symmetrically along the boundary inside $B_1(c_j)$ without intersecting each other or $B_\alpha(z)$ as in the left image. By continuity that can also be done for slightly bigger alphas. Let $\alpha(z)$ the biggest value of α for which that can be done. For the value of z in this example and the position of the balls $B_1(c_t)$, we have $\alpha(z) \approx 1.1$, and the intersections copied inside $B_1(c_j)$ are represented in the image at the right.

Let $\tau_j = \tau_j(\alpha, \ldots, \alpha)$. For every $z \in B_{\tau_j}(c_j)$ it only sees its own cluster and nothing of the rest. Let $v \in \mathbb{R}^m$, $\|v\| = 1$ and consider the partial

derivative with respect to t along the line $z = c_j + tv : t \in (-\tau_j, \tau_j)$.

$$\mathbb{E}P_i^{(\alpha,\dots,\alpha)}(c_j + tv) = \int_{B_1(c_j)} \alpha - d(x, c_j + tv)\, d\mu_j x$$
$$\frac{\partial}{\partial t}\mathbb{E}P_i^{(\alpha,\dots,\alpha)}(c_j + tv) = \begin{cases} > 0 & \text{if} \quad -\tau_j < t < 0 \\ = 0 & \text{if} \quad t = 0 \\ < 0 & \text{if} \quad 0 < t < \tau_j \end{cases} \tag{4.46}$$

Then $c_j = \operatorname{argmax}_{z \in B_{\tau_j}(c_j)} \mathbb{E}P^{(\alpha,\dots,\alpha)}(z)$. And because of the way α was chosen, since the measures μ_t are translations of the same rotationally symmetric measure, if $z \in B_1(c_j) \backslash B_{\tau_j}(c_j)$ we have

$$\mathbb{E}P_i^{(\alpha,\dots,\alpha)}(z) = \int_{B_1(c_j) \cap B_\alpha(z)} \alpha - d(x,z)\, d\mu_j x + \sum_{i \neq j} \int_{B_1(c_i) \cap B_\alpha(z)} \alpha - d(x,z)\, d\mu_i x$$
$$< \int_{B_1(c_j)} \alpha - d(x, c_j)\, d\mu_j x = \mathbb{E}P_i^{(\alpha,\dots,\alpha)}(c_j).$$

This proves the claim in Step 1. □

Lemma 4.2.5. *There exists some $v > 0$ with the property: if $\alpha_1, \dots, \alpha_k \in (\alpha - v, \alpha + v)$ then the maximum of $\mathbb{E}P_i^{(\alpha_1,\dots,\alpha_k)}(z)$ restricted to $B_1(c_j)$ is attained at $z = c_j$.*

Proof. By continuity of $\mathbb{E}P^{(\alpha_1,\dots,\alpha_k)}(z)$ with respect to the parameters $\alpha_1, \dots, \alpha_k$ given $\varepsilon > 0$ there exists $v > 0$ such that if $\alpha - v < \alpha_j < \alpha + v$ for all $j = 1, \dots, k$, then $\operatorname{argmax}_{z \in B_1(c_j)} \mathbb{E}P_i^{(\alpha_1,\dots,\alpha_k)}(z) \in B_\varepsilon(c_j)$. Let choose $\varepsilon > 0$ and $v > 0$ small enough such that it is also true that $\varepsilon < \tau_j(\alpha_1, \dots, \alpha_k)$ for all $\alpha_1 \dots, \alpha_k \in (\alpha - v, \alpha + v)$. Then the derivative computation (4.46) applies,

and can conclude that for all $\alpha_1, \ldots, \alpha_k \in (\alpha - \upsilon, \alpha + \upsilon)$

$$\text{argmax}_{z \in B_1(c_j)} \mathbb{E}P_i^{(\alpha_1,\ldots,\alpha_k)}(z) = c_j.$$

□

Lemma 4.2.6. *Let* $\alpha_1, \ldots, \alpha_k$ *be such that* $\text{argmax}_{z \in B_1(c_j)} \mathbb{E}P^{(\alpha_1,\ldots,\alpha_k)}(z) = c_j$. *Let also assume there exists some* $x_{j,i} \in B_\tau(c_j)$ *where* $\tau < \tau_j$. *Then the maximum of* $P^{(\alpha_1,\ldots,\alpha_k)}(x_{j,1}), \ldots, P^{(\alpha_1,\ldots,\alpha_k)}(x_{j,n})$ *is attained for an* $x_{j,s}$ *in* $B_{\tau_j}(c_j)$ *with probability at least* $\beta(n)$ *where* $\lim_n \beta(n) = 1$.

Proof. Let M such that $0 < P_i^{(\alpha_1,\ldots,\alpha_k)}(z) < M$. Then we use Hoeffding's inequality,

$$\Pr\left(|P^{(\alpha_1,\ldots,\alpha_k)}(z) - \mathbb{E}P^{(\alpha_1,\ldots,\alpha_k)}(z)| > r\right) < 2\exp\left(\frac{-2r^2}{nM^2}\right)$$

We know $\text{argmax}_{z \in B_1(c_j)} \mathbb{E}P^{(\alpha_1,\ldots,\alpha_k)}(z) = c_j$ then by continuity there exists $0 < \tau' < \tau_j$ such that

$$\inf_{z \in B_{\tau'}(c_j)} \mathbb{E}P^{(\alpha_1,\ldots,\alpha_k)}(z) \geqslant \sup_{z \in B_1(c_j) \setminus B_{\tau'}(c_j)} \mathbb{E}P^{(\alpha_1,\ldots,\alpha_k)}(z).$$

Without loss of generality say $\tau' = \tau_j$. Every point inside $B_{\tau_j}(c_j)$ sees exactly its own cluster, the function $\mathbb{E}P^{(\alpha_1,\ldots,\alpha_k)}(z)$ is rotationally symmetric since the measure is rotationally symmetric, and if we consider $z = c_j + te_1$ then it is increasing in t for $t \in (-\tau_j, 0)$ and decreasing for $t \in (0, \tau_j)$.

Let r and n satisfy

$$n\mathbb{E}P_i^{(\alpha_1,\ldots,\alpha_k)}(\tau_j e_1 + c_j) - r < n\mathbb{E}P_i^{(\alpha_1,\ldots,\alpha_k)}(\tau e_1 + c_j) + r \quad (\text{i.e. } r < Cn) \tag{4.47}$$

$$2\exp\left(\frac{-2r^2}{nM^2}\right) < 1 - \beta \quad (\text{i.e. } r > C'\sqrt{n}) \tag{4.48}$$

With high probability, the bigger $P^{(\alpha_1,\ldots,\alpha_k)}(z)$ is for z outside $B_{\tau_j}(c_j)$, the smaller the same function can be for $z \in B_\tau(c_j)$. In other words, if $x \in B_\tau(c_j)$ and $x' \in B_1(c_j) \backslash B_{\tau_j}(c_j)$

$$\Pr\left(|P^{(\alpha_A,\alpha_B)}(x) > P^{(\alpha_A,\alpha_B)}(x')|\right) > \beta.$$

This completes the proof of Theorem 4.2.3.

□

4.3 An integrality gap for the k-means LP relaxation

We now show that, in contrast to the LP relaxation for the k-medians clustering problem, the natural LP relaxation for k-means does not attain integral solutions for the stochastic ball model unless the separation between cluster centers exceeds $\Delta = 4$. This is a negative result because for separation $\Delta \geqslant 4$ the clustering problem becomes trivial. Every point is closer to points in its cluster than points in other clusters, therefore a simple thresholding algorithm will find the planted clusters.

Using the same notation than in the previous section, the LP relaxation for (k-means) is given by (k-means lp) below, whose dual LP is (k-means lp dual):

$$\begin{aligned} \underset{z\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad & \sum_{p,q\in P} d^2(p,q) z_{pq} && \text{(k-means lp)} \\ \text{subject to} \quad & \sum_{q\in P} z_{pq} = 1 \; \forall p \in P, \;\; z_{pq} \leqslant z_{pp} \; \forall p,q \in P, \\ & \sum_{p\in P} z_{pp} = k, \;\; z_{pq} \in [0,1] \end{aligned}$$

$$\begin{aligned} \underset{\substack{\alpha\in\mathbb{R}^n,\xi\in\mathbb{R} \\ \beta\in\mathbb{R}^{n\times n}}}{\text{maximize}} \quad & \sum_{p\in P} \alpha_p - k\xi && \text{(k-means lp dual)} \\ \text{subject to} \quad & \alpha_p \leqslant d^2(p,q) + \beta_{pq} \; \forall p,q \in P, \\ & \sum_{q\in P} \beta_{pq} = \xi \; \forall p \in P, \;\; \beta_{pq} \geqslant 0 \end{aligned}$$

In an integral solution to (k-means lp), the variable $z_{pq} = 1/|C|$ if p, q belong to the same cluster C in an optimal clustering, and $z_{pq} = 0$ otherwise. It is easy to see that such a solution satisfies all the constraints, and that the objective exactly measures the sum of average distances within every cluster. The following theorem shows the LP relaxation cannot recover the optimum k-means cluster solution if the distance between any two points in the same cluster is smaller than the distance between any two points in different clusters.

Theorem 4.3.1. *Given a set of points* $P = A_1 \cup \ldots \cup A_k$*, if the solution of* (k-means lp) *is integral and divides the set* P *in* k *clusters* $A_1, \ldots, A_k$ *then for all* p, q *in the same cluster* A_i *and* r *in a different cluster* A_j*,*

$$d(p,q) < d(p,r). \tag{4.49}$$

Proof. If the solution of (k-means lp) is integral and divides the set P in the clusters $A_1, \ldots, A_k$, complementary slackness tells us that

$$\alpha_p = d^2(p,q) + \beta_{pq} \quad \text{if } p,q \text{ are in the same cluster} \tag{4.50}$$

$$\beta_{pr} = 0 \quad \text{if } p,r \text{ are in different clusters} \tag{4.51}$$

if and only if α, β are corresponding optimal dual variables. Combining (k-means lp dual), (4.50) and (4.51), since $\beta_{pq} > 0$ we obtain that if p,q are in the same cluster and r is in a different cluster,

$$d^2(p,q) + \beta_{pq} = \alpha_p \leqslant d^2(p,r) \tag{4.52}$$

□

The result in Theorem 4.3.1 is tight in the sense of our distributional model. The following theorem shows separation $\Delta = 4$ is a threshold for cluster recovery via k-means LP.

Theorem 4.3.2. *Consider points drawn from the $(\mathcal{D}, \gamma, n)$-stochastic ball model. If n is sufficiently large, then the solution of* (k-means lp) *does not coincide with the planted clusters with high probability for $\Delta < 4$ and it does coincide for $\Delta > 4$.*

Proof. For $\Delta < 4$ the result in Theorem 4.3.1 implies that the solution of the LP will not be the planted clustering with high probability if enough points are provided.

For $\Delta > 4$ we show $z_{pq} = \begin{cases} 1/|C| & \text{if } p,q \text{ belong to the same cluster } C \\ 0 & \text{otherwise} \end{cases}$ is the solution of (k-means lp).

For α's and β's feasible for the dual problem we have $\sum_{q\in P}\beta_{pq} = \xi$ for all $p \in P$ implies $\sum_{p,q\in P}\beta_{pq}z_{pq} = k\xi$; we also have (as a consequence of (4.50) and the definition of z_{pq}) that $\alpha_p = \sum_{q\in P}(d^2(p,q) + \beta_{pq})z_{pq}$. Then for any dual feasible solution we have

$$\sum_{p,q\in P} d^2(p,q)z_{pq} = \sum_{p\in P}\alpha_p - k\xi.$$

Therefore, the existence of a feasible solution for the dual implies that our planted solution is optimal. Then it remains to show that there exists a feasible point for the dual. The solution is generically unique because no constraint in (k-means lp) is parallel to the objective function.

Existence of feasible solution of the dual

A feasible solution of the dual is $\{\alpha_p\}_{p\in P}$, $\{\beta_{pq}\}_{p,q\in P}$ such that (4.50), (4.51) are satisfied together with $\beta_{pq} \geqslant 0$ for all $p, q \in P$ and $\sum_{q\in P}\beta_{pq} = \xi$ for all $p \in P$. For $p \in P$ let C_p its cluster, $|C_p| = n$, then summing (4.50) in $q \in C_p$ we get

$$n\alpha_p = \sum_{q\in C_p} d^2(p,q) + \xi.$$

Let $\mathrm{avg}(p) := \frac{1}{n}\sum_{q\in C_p} d^2(p,q)$ and $\alpha_p := \mathrm{avg}(p) + \frac{\xi}{n}$, and define $m_{in}(p) := \max_{q\in C_p} d^2(p,q)$ and $m_{out}(p) := \min_{r\notin C_p} d^2(p,r)$. Assuming there exists a feasible point for the dual we know the solution for the LP is integral (i.e. our planted clustering) then we know (4.52) holds. In other words:

$$m_{in}(p) \leqslant \alpha_p \leqslant m_{out}(p) \text{ for all } p \in P$$

Equivalently,

$$m_{in}(p) - \text{avg}(p) \leqslant \frac{\xi}{n} \leqslant m_{out}(p) - \text{avg}(p) \text{ for all } p \in P \tag{4.53}$$

Then, a feasible solution for the dual problem exists if there exists ξ that satisfies (4.53) for all $p \in P$. A sufficient condition is:

$$\max_{r \in P} m_{in}(r) - \text{avg}(r) \leqslant \min_{s \in P} m_{out}(s) - \text{avg}(s)$$

Since this condition does not depend on the position of the cluster we can assume that the cluster C_r where the LHS is maximized is centered in 0. Let $f(r) = m_{in}(r) - \text{avg}(r) = \frac{1}{n}\sum_{l \in C_r} \|r - m_{in}(r)\|^2 - \|r - l\|^2$. In order to find its maximum consider

$$\frac{\partial f}{\partial r} = \frac{1}{n}\sum_{l \in C_r} 2(r - m_{in}(r)) - 2(r - l) = \frac{1}{n}\sum_{l \in C_r} -2m_{in}(r) \text{ since } C_r \text{ has mean } 0$$

But $m_{in}(r) \neq 0$ for all $r \in P$ since the center of the cluster cannot maximize the distance square (unless the trivial case where all the points in the cluster coincide with the center). Then f is maximized in the boundary of the unit ball. Then we need

$$4 - \min_{r \in \partial C} \text{avg}(r) \leqslant (\Delta - 2)^2 - \max_{s \in \partial C} \text{avg}(s)$$

which holds for $\Delta > 4$ with high probability when $n \to \infty$ since the points come from a rotationally symmetric distribution. □

Chapter 5

Efficiently certifying exact solutions

As mentioned in the introduction Lloyd's algorithm and its variants [7, 54] are fast but may converge to local minima of the k-means objective (e.g., see Figure 1.6 and section 5 of [8]). Furthermore, the output of Lloyd's algorithm does not indicate how far it is from optimal. In fact, most non-convex optimization methods fail to produce a certificate of global optimality. However, if a non-convex problem enjoys a convex relaxation, then solving the dual of this relaxation will produce a certificate of (approximate) optimality.

In the previous chapter we show that under the stochastic ball model, the convex relaxation (k-means sdp) is *tight* with high probability, that is, every solution to the relaxed problem (k-means sdp) identifies an optimal clustering. As such, in this high-probability event, one may solve the dual

This chapter is based on the publication:
Takayuki Iguchi, Dustin G. Mixon, Jesse Peterson, Soledad Villar. *Probably certifiably correct* k*-means clustering* Mathematical Programming, 2016 (to appear).
The author contributed in developing the main ideas of the paper, the mathematical proofs and numerical experiments.

program to produce a certificate of optimality. However, semidefinite programming (SDP) solvers are notoriously slow. For example, running MATLAB's built-in implementation of Lloyd's algorithm on 64 points in $\mathbb{R}^6$ will take about 0.001 seconds, whereas a CVX implementation [29] of the dual of (k-means sdp) for the same data takes about 20 seconds. Also, Lloyd's algorithm scales much better than SDP solvers, and so one should expect this runtime disparity to only increase with larger datasets. Overall, while the SDP relaxation theoretically produces a certificate in polynomial time (e.g., by an interior-point method [51]), it is far too slow to wait for in practice.

We combine the best of both worlds to obtain a new sort of algorithm, recently introduced by Bandeira in [9]:

Definition 5.0.1. Let **P** be an optimization problem that depends on some input, and let **D** denote a probability distribution over possible inputs. Then a **probably certifiably correct (PCC) algorithm** for $(\mathbf{P}, \mathbf{D})$ is an algorithm that on input $D \sim \mathbf{D}$ produces a global optimizer of **P** with high probability, and furthermore produces a certificate of having done so.

As mentioned in the introduction, the general technique to certify global optimality leverages several components simultaneously:

(i) A fast non-convex solver that produces the optimal solution with high probability (under some reasonable probability distribution of problem instances).

(ii) A convex relaxation that is tight with high probability (under the same distribution).

(iii) A fast method of computing a certificate of global optimality for the output of the non-convex solver in (i) by exploiting convex duality with the relaxation in (ii).

In the context of k-means, one might expect Lloyd's algorithm and the Peng–Wei SDP (k-means sdp) to be suitable choices for (i) and (ii), respectively. For (iii), one might adapt Bandeira's original method in [9] based on complementary slackness (see Figure 5.1 for an illustration). In this chapter, we provide a theoretical basis for each of these components in the context of k-means.

Note that when we proved tightness of the SDP in Section 4.1 we accomplished component (ii) in Bandeira's PCC technique, we tackle component (iii) next. For this, we recall Theorem 4.1.7, and express it in the following terms:

Theorem. *4.1.7. Take* X *a cluster projection matrix, and let* $P_{\Lambda^\perp}$ *denote the orthogonal projection onto the orthogonal complement of the span of* $\{1_{A_t}\}_{t=1}^k$. *Then there exists an explicit matrix* $Z = Z(D,X)$ *and scalar* $z = z(D,X)$ *such that* X *is a solution to the semidefinite relaxation* (k-means sdp) *if*

$$P_{\Lambda^\perp} Z P_{\Lambda^\perp} \preceq z P_{\Lambda^\perp}. \tag{5.1}$$

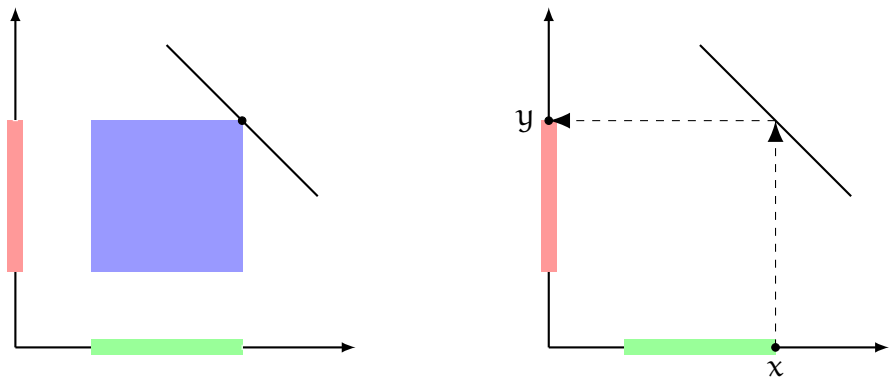


Figure 5.1: Complementary slackness and probably certifiably correct algorithms **(left)** Depiction of complementary slackness. The horizontal axis represents a vector space in which we consider a cone program (e.g., a linear or semidefinite program), and the feasibility region of this program is highlighted in green. The dual program concerns another vector space, which we represent with the vertical axis and feasibility region highlighted in red. The downward-sloping line represents all pairs of points (x, y) that satisfy complementary slackness. Recall that when strong duality is satisfied, we have that x is primal-optimal and y is dual-optimal if and only if x is primal feasible, y is dual feasible, and (x, y) satisfy complementary slackness. As such, the intersection between the blue Cartesian product and the complementary slackness line represents all pairs of optimizers. **(right)** Bandeira's probably certifiably correct technique [9]. Given a purported primal-optimizer x, we first check that x is primal-feasible. Next, we select y such that (x, y) satisfies complementary slackness. Finally, we check that y is dual-feasible. By complementary slackness, y is then a dual certificate of x's optimality in the primal program, which can be verified by comparing their values (a la strong duality).

Now we consider the matrix

$$A := \frac{z}{N} 1 1^\top + P_{\Lambda^\perp} Z P_{\Lambda^\perp}, \tag{5.2}$$

where z and Z come from Theorem 4.1.7. Since the all-ones vector 1 lies in the span of $\{1_{A_t}\}_{t=1}^k$, we have that 1 spans the unique leading eigenspace of A precisely when $P_{\Lambda^\perp} Z P_{\Lambda^\perp} \prec z P_{\Lambda^\perp}$, which in turn implies that X is a k-means optimal clustering by Theorem 4.1.7. As such, component (iii) can be accomplished by solving the following fundamental problem from linear algebra:

Problem. Given a symmetric matrix $A \in \mathbb{R}^{n \times n}$ and an eigenvector v of A, determine whether the span of v is the unique leading eigenspace, that is, the corresponding eigenvalue λ has multiplicity 1 and satisfies $|\lambda| > |\lambda'|$ for every other eigenvalue λ' of A.

Interestingly, this same problem appeared in Bandeira's original PCC theory [9], but it was left unresolved. In this chapter, we fill this gap by developing a so-called power iteration detector, which applies the power iteration to a random initialization on the unit sphere. Due to the randomness, the power iteration produces a test statistic that allows us to infer whether (A, v) satisfies the desired leading eigenspace condition. In Section 5.1, we pose this as a hypothesis test, and we estimate the associated error probabilities. In addition, we show how to leverage the structure of A defined by (5.2) and Theorem 5.1 to compute the matrix–vector multiplication Ax

for any given x in only $O(kmN)$ operations, thereby allowing the test statistic to be computed in linear time (up to the spectral gap of A and the desired confidence for the hypothesis test). Overall, the power iteration detector will deliver a highly confident inference on whether (A, v) satisfies the leading eigenspace condition, which in turn certifies the optimality of X up to the prescribed confidence level. Of course, one may remove the need for a confidence level by opting for deterministic spectral methods, but we do not know how to accomplish this in linear or even near-linear time.

5.1 A fast test for k-means optimality

In this section, we leverage the certificate (4.24) to test the optimality of a candidate k-means solution. We first show how to solve a more general problem from linear algebra, and then we apply our solution to devise a fast test for k-means optimality (as well as fast test for a related PCC algorithm).

5.1.1 Leading eigenvector hypothesis test

This subsection concerns Problem 5. To solve this problem, one might be inclined to apply the power method:

Proposition 5.1.1 (Theorem 8.2.1 in [28]). *Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix with eigenvalues $\{\lambda_i\}_{i=1}^n$ (counting multiplicities) satisfying*

$$|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|,$$

and with corresponding orthonormal eigenvectors $\{v_i\}_{i=1}^n$. Pick a unit-norm vec-

tor $q_0 \in \mathbb{R}^n$ *and consider the power iteration* $q_{j+1} := Aq_j/\|Aq_j\|_2$. *If* q_0 *is not orthogonal to* v_1, *then*

$$(v_1^\top q_j)^2 \geqslant 1 - \left((v_1^\top q_0)^{-2} - 1\right)\left(\frac{\lambda_2}{\lambda_1}\right)^{2j}.$$

Notice that the above convergence guarantee depends on the quality of the initialization q_0. To use this guarantee, draw q_0 at random from the unit sphere so that q_0 is not orthogonal to v_1 almost surely; one might then analyze the statistics of $v_1^\top q_0$ to produce statistics on the time required for convergence. The power method is typically used to find a leading eigenvector, but for our problem, we already have access to an eigenvector v, and we are tasked with determining whether v is the unique leading eigenvector. Intuitively, if you run the power method from a random initialization and it happens to converge to v, then this would have been a remarkable coincidence if v were not the unique leading eigenvector. Since we will only run finitely many iterations, how do we decide when we are sufficiently confident? The remainder of this subsection answers this question.

Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$ and a unit eigenvector v of A, consider the hypotheses

$$\begin{aligned} H_0&: \quad \text{span}(v) \text{ is not the unique leading eigenspace of } A, \\ H_1&: \quad \text{span}(v) \text{ is the unique leading eigenspace of } A. \end{aligned} \tag{5.3}$$

To test these hypotheses, pick a tolerance $\epsilon > 0$ and run the power iteration detector (see Algorithm 2). This detector terminates either by accepting H_0 or by rejecting H_0 and accepting H_1. We say the detector **fails to reject** H_0 if

it either accepts H_0 or fails to terminate. Before analyzing this detector, we consider the following definition:

Algorithm 2: Power iteration detector

Input: Symmetric matrix $A \in \mathbb{R}^{n\times n}$, unit eigenvector $v \in \mathbb{R}^n$, tolerance $\epsilon > 0$

Output: Decision of whether to accept H_0 or to reject H_0 and accept H_1 as given in (5.3)

$\lambda \leftarrow v^\top A v$
Draw q uniformly at random from the unit sphere in $\mathbb{R}^n$
while *no decision has been made* **do**
  if $|q^\top A q| > |\lambda|$ **then**
    Print `accept` H_0
  else if $(v^\top q)^2 \geqslant 1 - \epsilon$ **then**
    Print `reject` H_0 `and accept` H_1
  $q \leftarrow Aq/\|Aq\|_2$

Definition 5.1.1. Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$ and unit eigenvector v of A, put $\lambda = v^\top A v$, and let λ_1 denote a leading eigenvalue of A (i.e., $|\lambda_1| = \|A\|$). We say (A, v) is **degenerate** if

(a) the eigenvalue λ of A has multiplicity $\geqslant 2$,

(b) $-\lambda$ is an eigenvalue of A, or

(c) $-\lambda_1$ is an eigenvalue of A.

Theorem 5.1.2. *Consider the power iteration detector (Algorithm 2), let* q_j *denote* q *at the* j*th iteration (with* q_0 *being the initialization), and let* π_ϵ *denote the probability that* $(e_1^\top q_0)^2 < \epsilon$.

(i) (A, v) *is degenerate only if* H_0 *holds. If* (A, v) *is non-degenerate, then the power iteration detector terminates in finite time with probability* 1.

(ii) The power iteration detector incurs the following error rates:

$$\Pr\left(\textit{reject } H_0 \textit{ and accept } H_1 \;\middle|\; H_0 \right) \leqslant \pi_\epsilon,$$
$$\Pr\left(\textit{fail to reject } H_0 \;\middle|\; H_1 \right) = 0.$$

(iii) If H_1 *holds, then*

$$\min\left\{ j : (v^\top q_j)^2 > 1 - \epsilon \right\} \leqslant \frac{3\log(1/\epsilon)}{2\log|\lambda_1/\lambda_2|} + 1$$

with probability $\geqslant 1 - \pi_\epsilon$*, where* λ_2 *is the second largest eigenvalue (in absolute value).*

Proof. Denote the eigenvalues of A by $\{\lambda_i\}_{i=1}^n$ (counting multiplicities), ordered in such a way that $|\lambda_1| \geqslant \cdots \geqslant |\lambda_n|$, and consider the corresponding orthonormal eigenvectors $\{v_i\}_{i=1}^n$, where $v = v_p$ for some p.

For (i), first note that H_1 implies that (A, v) is non-degenerate, and so the contrapositive gives the first claim. Next, suppose (A, v) is non-degenerate. If H_1 holds, then $(v^\top q_j)^2 \to 1$ by Proposition 5.1.1 provided q_0 is not orthogonal to v, and so the power iteration detector terminates with probability 1. Otherwise, H_0 holds, and so the non-degeneracy of (A, v) implies that the eigenspace corresponding to λ_1 is the unique leading

eigenspace of A, and furthermore, $|\lambda_1| > |\lambda|$. Following the proof of Theorem 8.2.1 in [28], we also have

$$q_j^\top A q_j = \frac{q_0^\top A^{2j+1} q_0}{q_0^\top A^{2j} q_0} = \frac{\sum_{i=1}^n (v_i^\top q_j)^2 \lambda_i^{2j+1}}{\sum_{i=1}^n (v_i^\top q_j)^2 \lambda_i^{2j}}.$$

Putting $r := \min\{i : |\lambda_i| < |\lambda_1|\}$, then

$$\begin{aligned} |q_j^\top A q_j - \lambda_1| &= \left| \frac{\sum_{i=1}^n (v_i^\top q_j)^2 \lambda_i^{2j} (\lambda_i - \lambda_1)}{\sum_{i=1}^n (v_i^\top q_j)^2 \lambda_i^{2j}} \right| \\ &\leqslant \frac{|\lambda_1 - \lambda_n|}{\|P_{\lambda_1} q_0\|_2^2} \sum_{i=r}^n (v_i^\top q_j)^2 \left(\frac{\lambda_i}{\lambda_1}\right)^{2j} \\ &\leqslant |\lambda_1 - \lambda_n| \left(\frac{1 - \|P_{\lambda_1} q_0\|_2^2}{\|P_{\lambda_1} q_0\|_2^2}\right) \left(\frac{\lambda_r}{\lambda_1}\right)^{2j}, \end{aligned}$$

where P_{λ_1} denotes the orthogonal projection onto the eigenspace corresponding to λ_1. As such, $|q_j^\top A q_j| \to |\lambda_1| > |\lambda|$ provided $P_{\lambda_1} q_0 \neq 0$, and so the power iteration detector terminates with probability 1.

For (ii), we first consider the case of a false positive. Taking $v = v_p$ for $p \neq 1$, note that $(v^\top q_j)^2 > 1 - \epsilon$ implies

$$\epsilon > 1 - (v^\top q_j)^2 = \|q_j\|_2^2 - (v_p^\top q_j)^2 = \sum_{\substack{i=1 \\ i \neq p}}^n (v_i^\top q_j)^2 \geqslant (v_1^\top q_j)^2.$$

Also, since $\|Ax\|_2 \leqslant |\lambda_1| \|x\|_2$ for all $x \in \mathbb{R}^n$, we have that $(v_1^\top q_j)^2$ monotonically increases with j:

$$(v_1^\top q_{j+1})^2 = \left(v_1^\top \frac{A q_j}{\|A q_j\|_2}\right)^2 = \frac{(\lambda_1 v_1^\top q_j)^2}{\|A q_j\|_2^2} \geqslant \frac{(v_1^\top q_j)^2}{\|q_j\|^2} = (v_1^\top q_j)^2.$$

As such, $\epsilon > (v_1^\top q_j)^2 \geqslant (v_1^\top q_0)^2$. Overall, when H_0 holds, the power iteration detector rejects H_0 only if q_0 is initialized poorly, i.e., $(v_1^\top q_0)^2 < \epsilon$, which

occurs with probability π_ϵ (since q_0 has a rotation-invariant probability distribution). For the false negative error rate, note that Proposition 5.1.1 gives that H_1 implies convergence $(v^\top q_j)^2 \to 1$ provided q_0 is not orthogonal to v, i.e., with probability 1.

For (iii), we want j such that $(v^\top q_j)^2 > 1 - \epsilon$. By Proposition 5.1.1, it suffices to have

$$\left((v_1^\top q_0)^{-2} - 1\right)\left(\frac{\lambda_2}{\lambda_1}\right)^{2j} < \epsilon.$$

In the event that $(v_1^\top q_0)^2 \geqslant \epsilon$ (which has probability $1 - \pi_\epsilon$), it further suffices to have

$$\epsilon^{-2}\left(\frac{\lambda_2}{\lambda_1}\right)^{2j} < \epsilon.$$

Taking logs and rearranging then gives the result. □

To estimate ϵ and π_ϵ, first note that q_0 has a rotation-invariant probability distribution, and so linearity of expectation gives

$$\mathbb{E}\left[(e_1^\top q_0)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(e_i^\top q_0)^2\right] = \frac{1}{n}\mathbb{E}\|q_0\|_2^2 = \frac{1}{n}.$$

Thus, in order to make π_ϵ small, we should expect to have $\epsilon \ll 1/n$. The following lemma gives that such choices of ϵ suffice for π_ϵ to be small:

Lemma 5.1.3. *If* $\epsilon \geqslant n^{-1}e^{-2n}$, *then* $\pi_\epsilon \leqslant 3\sqrt{n\epsilon}$.

Proof. First, observe that $(e_1^\top q_0)^2$ is equal in distribution to Z^2/Q, where Z has standard normal distribution and Q has chi-squared distribution with

n degrees of freedom (Z and Q are independent). The probability density function of Z has a maximal value of $1/\sqrt{2\pi}$ at zero, and so

$$\Pr\left(Z^2 < a\right) \leqslant \sqrt{\frac{2a}{\pi}}.$$

Also, Lemma 1 in [39] gives

$$\Pr\left(Q \geqslant n + 2\sqrt{nx} + 2x\right) \leqslant e^{-x} \qquad \forall x > 0.$$

Therefore, picking $a = 5n\epsilon$ and $x = n$, the union bound gives

$$\begin{aligned}\Pr\left((e_1^\top q_0)^2 < \epsilon\right) = \Pr\left(\frac{Z^2}{Q} < \epsilon\right) &\leqslant \Pr\left(Z^2 < 5n\epsilon\right) + \Pr\left(Q > 5n\right) \\ &\leqslant \sqrt{\frac{10n\epsilon}{\pi}} + e^{-n} \leqslant 3\sqrt{n\epsilon}.\end{aligned}$$

□

Overall, if we take $\epsilon = n^{-(2c+1)}$ for $c > 0$, then if H_0 is true, our detector will produce a false positive with probability $O(n^{-c})$. On the other hand, if H_1 is true, then with probability $1 - O(n^{-c})$, our detector will reject H_0 after $O_\delta(c \log n)$ power iterations, provided $|\lambda_2| \leqslant (1-\delta)|\lambda_1|$.

5.1.2 Testing optimality with the power iteration detector

In this subsection, we leverage the power iteration detector to test k-means optimality. Note that the sufficient condition (4.24) holds if and only if $v := \frac{1}{\sqrt{N}}1$ is a leading eigenvector of the matrix

$$A := \frac{z}{N}11^\top + P_{\Lambda^\perp}(B - M)P_{\Lambda^\perp} = \frac{z}{N}11^\top + P_{\Lambda^\perp}(B - D)P_{\Lambda^\perp}. \tag{5.4}$$

(The second equality follows from distributing the $P_{\Lambda^\perp}$'s and recalling the definition of M in (4.20).) As such, it suffices that (A, v) satisfy H_1 in (5.3).

Overall, given a collection of points $\{x_i\}_{i=1}^N \subseteq \mathbb{R}^m$ and a proposed partition $A_1 \sqcup \cdots \sqcup A_k = \{1, \ldots, N\}$, we can produce the corresponding matrix A (defined above) and then run the power iteration detector of the previous subsection to test (4.24). In particular, a positive test with tolerance ϵ will yield $\geqslant 1 - \pi_\epsilon$ confidence that the proposed partition is optimal under the k-means objective. Furthermore, as we detail below, the matrix–vector products computed in the power iteration detector have a computationally cheap implementation.

Given an $m \times n_a$ matrix $\Phi_a = [x_{a,1} \cdots x_{a,n_a}]$ for each $a \in \{1, \ldots, k\}$, we follow the following procedure to implement the corresponding function $x \mapsto Ax$ as defined in (5.4):

STEPS IN COMPUTATION OF $x \mapsto Ax$. *cost in operations*

1: Compute $\nu_a \in \mathbb{R}^{n_a}$ such that $(\nu_a)_i = \|x_{a,i}\|_2^2$ for every $a \in \{1, \ldots, k\}$.
Let $\nu \in \mathbb{R}^N$ denote the vector whose ath block is ν_a. $O(mN)$

2: Define the function $(a, b, x) \mapsto D^{(a,b)}x$ such that
$D^{(a,b)} = \nu_a 1^\top - 2\Phi_a^\top \Phi_b + 1\nu_b^\top$. $O(m(n_a + n_b))$

3: Define the function $x \mapsto Dx$ such that $D = \nu 1^\top - 2\Phi^\top \Phi + 1\nu^\top$,
where $\Phi = [\Phi_1 \cdots \Phi_k]$. $O(mN)$

4: Compute $\mu_a = \frac{1}{2}(\frac{1}{n_a^2} 1 1^\top - \frac{2}{n_a} I) D^{(a,a)} 1$ for every $a \in \{1, \ldots, k\}$. $O(mN)$

5: Define the function $(a, b, x) \mapsto M^{(a,b)}x$ such that
$M^{(a,b)} = D^{(a,b)} + \mu_a 1^\top + 1\mu_b^\top$. $O(m(n_a + n_b))$

6: Compute $z = \min_{a \neq b} \frac{2n_a}{n_a + n_b} \min(M^{(a,b)}1)$. $O(kmN)$

7: Compute $u_{(a,b)} = M^{(a,b)}1 - z\frac{n_a+n_b}{2n_a}1$ for every $a, b \in \{1, \ldots, k\}$, $a \neq b$. $O(kmN)$

8: Compute $\rho_{(a,b)} = u_{(a,b)}^\top 1$ for every $a, b \in \{1, \ldots, k\}$, $a \neq b$. $O(kN)$

9: Define the function $x \mapsto Bx$ such that the ath block of the output is given by $(Bx)_a = \sum_{\substack{b=1 \\ b \neq a}}^{k} \frac{u_{(a,b)} u_{(b,a)}^\top x_b}{\rho_{(b,a)}}$. $O(kmN)$

10: Define the function $x \mapsto P_{\Lambda^\perp} x$ such that $P_{\Lambda^\perp} = I - \sum_{a=1}^{k} \frac{1}{n_a} 1_a 1_a^\top$. $O(N)$

11: Define the function $x \mapsto Ax$ such that $A = \frac{z}{N} 11^\top + P_{\Lambda^\perp}(B - D)P_{\Lambda^\perp}$. $O(kmN)$

Overall, after $O(kmN)$ operations of preprocessing, one may compute the function $x \mapsto Ax$ for any given x in $O(kmN)$ operations. (Observe that this is the same complexity as each iteration of Lloyd's algorithm.)

At this point, we take a short aside to illustrate the utility of the power iteration detector beyond k-means clustering. The original problem for which a PCC algorithm was developed was community recovery under the **stochastic block model** [9]. For this random graph, there are two communities of vertices, each of size $n/2$, and edges are drawn independently at random with probability p if the pair of vertices belong to the same community, and with probability $q < p$ if they come from different communities. Given the random edges, the maximum likelihood estimator for the

communities is given by the vertex partition of two sets of size $n/2$ with the minimum cut. Given a partition of the vertices, let X denote the corresponding $n \times n$ matrix of ± 1s such that $X_{ij} = 1$ precisely when i and j belong to the same community. Given the adjacency matrix A of the random graph, one may express the cut of a partition X in terms of $\mathrm{Tr}(AX)$. Furthermore, X satisfies the convex constraints $X_{ii} = 1$ and $X \succeq 0$, and so one may relax to these constraints to obtain a semidefinite program and hope that the relaxation is typically tight over a large region of (p, q). Amazingly, this relaxation is typically tight precisely over the region of (p, q) for which community recovery is information-theoretically possible [2].

Given A, let $B := 2A - 11^\top + I$, and given a vector $x \in \mathbb{R}^n$, define the corresponding $n \times n$ diagonal matrix D_x by $(D_x)_{ii} := x_i \sum_{j=1}^{n} B_{ij} x_j$. In [9], Bandeira observes that, given a partition matrix X by some means (such as the fast algorithm provided in [3]), then $X = xx^\top$ is SDP-optimal if both $x^\top 1 = 0$ and the second smallest eigenvalue of $D_x - B$ is strictly positive, meaning the partition gives the maximum likelihood estimator for the communities. However, as Bandeira notes, the computational bottleneck here is estimating the second smallest eigenvalue of $D_x - B$, and he suggests that a randomized power method–like algorithm might suffice, but leaves the investigation for future research.

Here, we show how the power iteration detector fills this void in the theory. First, we note that in the interesting regime of (p, q), the number of nonzero entries in A is $O(n \log n)$ with high probability [2]. As such,

the function $x \mapsto Bx$ can exploit this sparsity to take only $O(n \log n)$ operations. This in turn allows for the computation of the diagonal of D_x to cost $O(n \log n)$ operations. Next, note that

$$\begin{aligned} \|D_x - B\| &\leqslant \|D_x\| + \|2A - 11^\top\| + \|I\| \\ &\leqslant \|D_x\| + \|2A - 11^\top\|_F + 1 = \max_i |(D_x)_{ii}| + n + 1 =: \lambda, \end{aligned}$$

and that λ can be computed in $O(n)$ operations after computing the diagonal of D_x. Also, it takes $O(n)$ operations to verify $x^\top 1 = 0$. Assuming $x^\top 1 = 0$, then the second smallest eigenvalue of $D_x - B$ is strictly positive if and only if x spans the unique leading eigenspace of $\lambda I - D_x + B$. Thus, one may test this condition using the power iteration detector, and furthermore, each iteration will take only $O(n \log n)$ operations, thanks to the sparsity of A.

Chapter 6

Approximation guarantees for relax and round algorithms

The results we present in this chapter appear in [49]. That paper has two main results. First, it presents a relax-and-round algorithm for k-means clustering that well-approximates the centers of sufficiently separated subgaussians. Second, it provides a conditional result on the minimum separation necessary for Gaussian center approximation by k-means clustering. The first result establishes that the k-means SDP (k-means sdp) performs well with noisy data (despite not being tight), and the second result helps to illustrate how sharp our analysis is.

In this thesis we only include the first result of the paper, regarding the approximation guarantees. We refer the reader to [49] for the conditional lower bound for learning mixtures of Gaussians using k-means. At the end

This chapter is based on the publication:
Dustin G. Mixon, Soledad Villar, Rachel Ward. *Clustering subgaussian mixtures by semidefinite programming.* Information and Inference: A Journal of the IMA, 2017 (to appear).
The author contributed in developing the main ideas of the paper, the mathematical proofs and numerical experiments.

of this chapter we apply our algorithm to the MNIST dataset [40]. See [66] for an interactive visualization of the numerical experiment.

6.1 The relax-and-round algorithm

As we mentioned earlier in the introduction and we further explain in Section 6.4, our relax-and-round algorithm relies on the interpretation of the matrix X as a relaxation of projection, and the matrix PX as a denoised version of the points P. The algorithm we propose consists of the following steps:

Relax-and-round k-means clustering procedure.
Given and $m \times N$ data matrix $P = [x_1 \cdots x_N]$, do:

(i) Compute distance-squared matrix D defined by $D_{ij} = \|x_i - x_j\|_2^2$.

(ii) Solve (k-means sdp), resulting in optimizer X.

(iii) Cluster the columns of the denoised data matrix PX.

For step (iii), we find there tends to be k vectors that appear as columns in PX with particularly high frequency, and so we are inclined to use these as estimators for the k-mean-optimal centroids (see Figures 1.7 and 6.1, for example). Running Lloyd's algorithm for step (iii) also works well in practice. To obtain theoretical guarantees, we instead find the k columns of PX

for which the unit balls of a certain radius centered at these points in $\mathbb{R}^m$ contain the most columns of PX (see Theorem 6.5.1 for more details).

6.2 Performance guarantee for the k-means SDP

Our relax-and-round performance guarantee consists of three steps.

Step 1: Approximation. We adapt an approach used by Guédon and Vershynin to provide approximation guarantees for a certain semidefinite program under the stochastic block model for graph clustering [32].

Given the points $x_{t,1}, \ldots, x_{t,n_t}$ drawn independently from $\mathcal{D}_t$, consider the squared-distance matrix D and the corresponding minimizer X_D of the SDP (k-means sdp). We first construct a "reference" matrix R such that the SDP (k-means sdp) is tight when $D = R$ with optimizer X_R. To this end, take $\Delta_{ab} := \|\gamma_a - \gamma_b\|_2$, let X_D denote the minimizer of (k-means sdp), and let X_R denote the minimizer of (k-means sdp) when D is replaced by the reference matrix R defined as

$$(R_{ab})_{ij} := \xi + \Delta_{ab}^2/2 + \max\left\{0, \Delta_{ab}^2/2 + 2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle\right\} \tag{6.1}$$

where $r_{t,i} := x_{t,i} - \gamma_t$, and $\xi > 0$ is a parameter to be chosen later. Indeed, this choice of reference is quite technical, as an artifact of the entries in D being statistically dependent. Despite its lack of beauty, our choice of reference enjoys the following important property:

Lemma 6.2.1. *Let $1_a \in \mathbb{R}^N$ denote the indicator function for the indices i corresponding to points x_i drawn from the ath subgaussian. If $\gamma_a \neq \gamma_b$ whenever $a \neq b$,*

then $X_R = \sum_{t=1}^{k} (1/n_t) 1_t 1_t^\top$.

Proof. Let X be feasible for the the SDP (k-means sdp). Replacing D with R in (k-means sdp), we may use the SDP constraints $X1 = 1$ and $X \geqslant 0$ to obtain the bound

$$\mathrm{Tr}(RX) = \sum_{i=1}^{N} \sum_{j=1}^{N} R_{ij} X_{ij} \geqslant \sum_{i=1}^{N} \sum_{j=1}^{N} \xi X_{ij} = \sum_{i=1}^{N} \xi \sum_{j=1}^{N} X_{ij} = N\xi = \mathrm{Tr}(RX_R)$$

Furthermore, since $\gamma_a \neq \gamma_b$ whenever $a \neq b$, and since $X \geqslant 0$, we have that equality occurs precisely for the X such that $(X_{ab})_{ij}$ equals zero whenever $a \neq b$. The other constraints on X then force X_R to have the claimed form (i.e., X_R is the unique minimizer). □

Now that we know that X_R is the planted solution, it remains to demonstrate regularity in the sense that $X_D \approx X_R$ provided the subgaussian centers are sufficiently separated. For this, we use the following scheme:

- If $\langle R, X_D \rangle \approx \langle R, X_R \rangle$ then $\|X_D - X_R\|_F^2$ is small (Lemma 6.3.1).
- If $D \approx R$ (in some specific sense) then $\langle R, X_D \rangle \approx \langle R, X_R \rangle$ (Lemmas 6.3.2 and 6.3.3).
- If the centers are separated by $O(k\sigma_{max})$, then $D \approx R$.

What follows is the result of this analysis:

Theorem 6.2.2. *Given* $x_1, \ldots, x_N$ *points drawn independently from a mixture of* k *subgaussian distributions in* $\mathbb{R}^m$. *Say that the subgaussian* a, *for* $1 \leqslant a \leqslant k$

has center γ_a*, maximum covariance* σ_a^2*, and* n_a *points have been drawn from it. Let* $n_{\max} := \max_{1 \leqslant a \leqslant k} n_a$ *and similarly* $n_{\min}$ *and* $\sigma_{\max}$*. Let* X_D *the minimizer of* (k-means sdp) *for* D *such that* $D_{ij} = \|x_i - x_j\|^2$ *and* X_R *the minimizer of* (k-means sdp) *for* $D = R$ *the reference matrix defined in* (6.1) *(which coincides with the planted clusters as a consequence of Lemma 6.2.1). Fix* $\epsilon, \eta > 0$*. There exist universal constants* C, c_1, c_2, c_3 *such that if*

$$\alpha = n_{\max}/n_{\min} \lesssim k \lesssim m \quad \textit{and} \quad N > \max\{c_1 m, c_2 \log(2/\eta), \log(c_3/\eta)\},$$

then $\|X_D - X_R\|_F^2 \leqslant \epsilon$ *with probability* $\geqslant 1 - 2\eta$ *provided*

$$\Delta_{\min}^2 \geqslant \frac{C}{\epsilon} k^2 \alpha \sigma_{\max}^2$$

where $\Delta_{\min} = \min_{a \neq b} \|\gamma_a - \gamma_b\|_2$ *is the minimal cluster center separation.*

See Section 6.3 for the proof. Note that if we remove the assumption $\alpha \lesssim k \lesssim m$, we obtain the result $\Delta_{\min}^2 \geqslant \frac{C}{\epsilon}(\min\{k, m\} + \alpha) k \alpha \sigma_{\max}^2$.

Step 2: Denoising. Suppose we solve the SDP (k-means sdp) for an instance of the subgaussian mixture model where $\Delta_{\min}$ is sufficiently large. Then Theorem 6.2.2 ensures that the solution X_D is close to the ground truth. At this point, it remains to convert X_D into an estimate for the centers $\{\gamma_t\}_{t \in [k]}$. Let P denote the $m \times N$ matrix whose (a, i)th column is $x_{a,i}$. Then PX_R is an $m \times N$ matrix whose (a, i)th column is $\tilde{\gamma}_a$, the centroid of the ath cluster, which converges to γ_a as $N \to \infty$, (and does not change when i varies, for a fixed a), and so one might expect PX_D to have its columns be

close to the γ_t's. In fact, we can interpret the columns of PX_D as a denoised version of the points (see Figure 1.7).

To illustrate this idea, assume the points $\{x_{a,i}\}_{i\in[n]}$ come from $\mathcal{N}(\gamma_a, \sigma^2 I_m)$ in $\mathbb{R}^m$ for each $a \in [k]$. Then we have

$$\mathbb{E}\left[\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n}\|x_{a,i}-\gamma_a\|_2^2\right] = m\sigma^2. \tag{6.2}$$

Letting $c_{a,i}$ denote the (a,i)th column of PX_D (i.e., the ith estimate of γ_a), in Section 6.4 we obtain the following denoising result:

Corollary 6.2.3. *If* $k\sigma \lesssim \Delta_{\min} \leqslant \Delta_{\max} \lesssim K\sigma$, *then*

$$\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n}\|c_{a,i}-\tilde{\gamma}_a\|_2^2 \lesssim K^2\sigma^2$$

with high probability as $n \to \infty$.

Note that Corollary 6.2.3 guarantees denoising in the regime $K \ll \sqrt{m}$. This is a corollary of a more technical result (Theorem 6.4.1), which guarantees denoising for certain configurations of subgaussians (e.g., when the γ_t's are vertices of a regular simplex) in the regime $k \ll m$.

At this point, we comment that one might expect this level of denoising from principal component analysis (PCA) when the mixture of subgaussians is sufficiently nice. To see this, suppose we have spherical Gaussians of equal entrywise variance σ^2 centered at vertices of a regular simplex. Then in the large-sample limit, we expect PCA to approach the $(k-1)$-dimensional affine subspace that contains the k centers. Projecting onto this

affine subspace will not change the variance of any Gaussian in any of the principal components, and so one expects the mean squared deviation of the projected points from their respective Gaussian centers to be $(k-1)\sigma^2$.

By contrast, we find that in practice, the SDP denoises substantially more than PCA does. For example, Figures 1.7 and 6.1 illustrate cases in which PCA would not change the data, since the data already lies in $(k-1)$-dimensional space, and yet the SDP considerably enhances the signal. In fact, we observe empirically that the matrix X_D has low rank and that PX_D has repeated columns. This doesn't come as a complete surprise, considering SDP optimizers are known to exhibit low rank [58, 11, 55]. Still, we observe that the optimizer tends to have rank $O(k)$ when clustering points from the mixture model. This is not predicted by existing bounds, and we did not leverage this feature in our analysis, though it certainly warrants further investigation.

Step 3: Rounding. In Section 6.5, we present a rounding scheme that provides a clustering of the original data from the denoised results of the SDP (Theorem 6.5.1). In general, the cost of rounding is a factor of k in the average squared deviation of our estimates. Under the same hypothesis as Corollary 6.2.3, we have that there exists a permutation π on $\{1,\dots,k\}$ such that

$$\frac{1}{k}\sum_{i=1}^{k}\|\nu_i-\tilde{\gamma}_{\pi(i)}\|_2^2 \lesssim kK^2\sigma^2, \tag{6.3}$$

where ν_i is what our algorithm chooses as the ith center estimate. Much like the denoising portion, we also have a more technical result that allows one

to replace the right-hand side of (6.3) with $k^2\sigma^2$ for sufficiently nice configurations of subgaussians. As such, we can estimate Gaussian centers with mean squared error $O(k^2\sigma^2)$ provided the centers have pairwise distance $\Omega(k\sigma)$. In the next section, we indicate that model order–dependence cannot be completely removed when using k-means to estimate the centers.

Before concluding this section, we want to clarify the nature of our approximation guarantee (6.3). Since centroids correspond to a partition of Euclidean space, our guarantee says something about how "close" our k-means partition is to the "true" partition. By contrast, the usual approximation guarantees for relax-and-round algorithms compare values of the objective function (e.g., the k-means value of the algorithm's output is within a factor of 2 of minimum). Also, the latter sort of optimal value–based approximation guarantee cannot be used to produce the sort of optimizer-based guarantee we want. To illustrate this, imagine a relax-and-round algorithm for k-means that produces a near-optimal partition with $k = 2$ for data coming from a single spherical Gaussian. We expect every subspace of co-dimension 1 to separate the data into a near-optimal partition, but the partitions are very different from each other when the dimension $m \geqslant 2$, and so a guarantee of the form (6.3) will not hold.

6.3 Proof of Theorem 6.2.2

By the following lemma, it suffices to bound $\mathrm{Tr}(R(X_D - X_R))$:

Lemma 6.3.1. $\|X_D - X_R\|_F^2 \leqslant \dfrac{5}{n_{\min}\Delta_{\min}^2} \operatorname{Tr}(R(X_D - X_R))$.

Proof. First, by Lemma 6.2.1, we have $\|X_R\|_F^2 = k$. We also claim that $\|X_D\|_F^2 \leqslant k$. To see this, first note that $X_D 1 = 1$ and $X_D \geqslant 0$, and so the ith entry of $X_D \nu$ can be interpreted as a convex combination of the entries of ν. Let ν be an eigenvector of X_D with eigenvalue μ, and let i index the largest entry of ν (this entry is positive without loss of generality). Then $\mu\nu_i = (X_D\nu)_i \leqslant \nu_i$, implying that $\mu \leqslant 1$. Since the eigenvalues of X_D lie in $[0,1]$, we may conclude that $\|X_D\|_F^2 \leqslant \operatorname{Tr}(X_D) = k$. As such,

$$\begin{aligned}
\|X_D - X_R\|_F^2 &= \|X_D\|_F^2 + \|X_R\|_F^2 - 2\operatorname{Tr}(X_D X_R) \\
&\leqslant 2k - 2\operatorname{Tr}(X_D X_R) \\
&= 2k + 2\operatorname{Tr}((X_R - X_D)X_R) - 2\|X_R\|_F^2 \\
&= 2\operatorname{Tr}((X_R - X_D)X_R). \qquad (6.4)
\end{aligned}$$

We will bound (6.4) in two different ways, and a convex combination of these bounds will give the result. For both bounds, we let Ω denote the indices in the diagonal blocks, and Ω^c the indices in the off-diagonal blocks, and $\Omega_t \subset \Omega$ denote the indices in the diagonal block for the cluster t. In particular, A_Ω denotes the matrix that equals A on the diagonal blocks and is zero on the off-diagonal blocks. For the first bound, we use that $R_\Omega = \xi(11^\top)_\Omega$, and that $(X_R - X_D)_\Omega(11^T)_\Omega$ has non-negative entries (since both X_R and X_D have non-negative entries, $X_R 1 = X_D 1 = 1$, and $X_R = (X_R)_\Omega$).

Recalling that $R_\Omega = \xi$, we have

$$\begin{aligned} 2\operatorname{Tr}((X_R - X_D)X_R) &= \sum_{t=1}^{k} 2\operatorname{Tr}\left((X_R - X_D)(11^\top)_{\Omega_t}\frac{1}{n_t}\right) \\ &\geqslant \frac{2}{n_{\max}}\operatorname{Tr}\left((X_R - X_D)(11^\top)_\Omega\right) \\ &= -\frac{2}{\xi n_{\max}}\operatorname{Tr}\left((X_D - X_R)R_\Omega\right) \end{aligned} \tag{6.5}$$

For the second bound, we first write

$$n_{\min}X_R = 11^\top - (11^\top)_{\Omega^c} - \sum_{t=1}^{k}\left(1 - \frac{n_{\min}}{n_t}\right)(11^\top)_{\Omega_t}.$$

Since $X_R 1 = 1 = X_D 1$, we then have

$$\begin{aligned} &2\operatorname{Tr}((X_R - X_D)X_R) \\ &\quad= \frac{2}{n_{\min}}\operatorname{Tr}\left((X_D - X_R)\left((11^\top)_{\Omega^c} + \sum_{t=1}^{k}\left(1 - \frac{n_{\min}}{n_t}\right)(11^\top)_{\Omega_t} - 11^\top\right)\right) \\ &\quad= \frac{2}{n_{\min}}\operatorname{Tr}\left((X_D - X_R)\left((11^\top)_{\Omega^c} + \sum_{t=1}^{k}\left(1 - \frac{n_{\min}}{n_t}\right)(11^\top)_{\Omega_t}\right)\right) \\ &\quad\leqslant \frac{2}{n_{\min}}\operatorname{Tr}((X_D - X_R)(11^\top)_{\Omega^c}) \\ &\quad= \frac{2}{n_{\min}}\operatorname{Tr}(X_D(11^\top)_{\Omega^c}), \end{aligned}$$

where the last and second-to-last steps use that $(X_R)_{\Omega^c} = 0$. Next, $X_D \geqslant 0$ and $R_{\Omega^c} \geqslant (\xi + \Delta_{\min}^2/2)(11^\top)_{\Omega^c}$, and so we may continue:

$$\begin{aligned} 2\operatorname{Tr}((X_R - X_D)X_R) &\leqslant \frac{2}{n_{\min}(\xi + \Delta_{\min}^2/2)}\operatorname{Tr}(X_D R_{\Omega^c}) \\ &= \frac{2}{n_{\min}(\xi + \Delta_{\min}^2/2)}\operatorname{Tr}((X_D - X_R)R_{\Omega^c}), \end{aligned} \tag{6.6}$$

where again, the last step uses the fact that $(X_R)_{\Omega^c} = 0$. At this point, we have bounds of the form $x \geqslant ay_1$ with $a < 0$ and $x \leqslant by_2$ with $b > 0$ (explicitly, (6.5) and (6.6)), and we seek a bound of the form $x \leqslant c(y_1 + y_2)$. As such, we take the convex combination for a, b such that $a^{-1}/(a^{-1} + b^{-1}) < 0$ and $b^{-1}/(a^{-1} + b^{-1}) > 0$

$$x \leqslant \frac{a^{-1}}{a^{-1} + b^{-1}} ay_1 + \frac{b^{-1}}{a^{-1} + b^{-1}} by_2 = \frac{1}{a^{-1} + b^{-1}}(y_1 + y_2).$$

Taking $a = -2/(\xi n_{max})$ and $b = 2/(n_{min}(\xi + \Delta_{min}^2/2))$ and combining with (6.4) then gives

$$\begin{aligned} \|X_D - X_R\|_F^2 &\leqslant 2\,\mathrm{Tr}((X_R - X_D)X_R) \\ &\leqslant \left(\frac{\xi}{2}(n_{min} - n_{max}) + \frac{n_{min}}{4}\Delta_{min}^2\right)^{-1} \mathrm{Tr}((X_D - X_R)(R_\Omega + R_{\Omega^c})), \end{aligned}$$

choosing $\xi > 0$ sufficiently small and simplifying yields the result. □

We will bound $\mathrm{Tr}(R(X_D - X_R))$ in terms of the following: For each $N \times N$ real symmetric matrix M, let $\mathcal{F}(M)$ denote the value of the following program:

$$\begin{aligned} \mathcal{F}(M) = \text{ maximum } & \quad |\mathrm{Tr}(MX)| \\ \text{subject to } & \quad \mathrm{Tr}(X) = k,\ X1 = 1,\ X \geqslant 0,\ X \succeq 0 \end{aligned} \tag{6.7}$$

Lemma 6.3.2. *Let* $\widetilde{D} := P_{1^\perp} D P_{1^\perp}$ *and* $\tilde{R} := P_{1^\perp} R P_{1^\perp}$. *Then*

$$\mathrm{Tr}(R(X_D - X_R)) \leqslant 2\mathcal{F}(\tilde{D} - \tilde{R}).$$

Proof. Since X_D and X_R are both feasible in (6.7), we have

$$-\mathrm{Tr}(\widetilde{D}X_D) + \mathrm{Tr}(\widetilde{R}X_D) \leqslant |\mathrm{Tr}((\widetilde{D}-\widetilde{R})X_D)| \leqslant \mathcal{F}(\widetilde{D}-\widetilde{R}),$$
$$\mathrm{Tr}(\widetilde{D}X_R) - \mathrm{Tr}(\widetilde{R}X_R) \leqslant |\mathrm{Tr}((\widetilde{D}-\widetilde{R})X_R)| \leqslant \mathcal{F}(\widetilde{D}-\widetilde{R}),$$

and adding followed by reverse triangle inequality gives

$$2\mathcal{F}(\widetilde{D}-\widetilde{R}) \geqslant \Big(\mathrm{Tr}(\widetilde{D}X_R) - \mathrm{Tr}(\widetilde{D}X_D)\Big) + \Big(\mathrm{Tr}(\widetilde{R}X_D) - \mathrm{Tr}(\widetilde{R}X_R)\Big). \tag{6.8}$$

Write $X_{\widetilde{D}} := P_{1^\perp} X_D P_{1^\perp}$. Note that $X_D 1 = (X_D)^T 1$ implies $X_D = X_{\widetilde{D}} + (1/N)11^\top$, and so

$$\mathrm{Tr}(\widetilde{D}X_D) = \mathrm{Tr}(DX_{\widetilde{D}}) = \mathrm{Tr}\Big(D\big(X_D - (1/N)11^\top\big)\Big) = \mathrm{Tr}(DX_D) + \frac{1}{N}1^\top D1.$$

Similarly, $\mathrm{Tr}(\widetilde{D}X_R) = \mathrm{Tr}(DX_R) + \frac{1}{N}1^\top D1$, and so

$$\mathrm{Tr}(\widetilde{D}X_R) - \mathrm{Tr}(\widetilde{D}X_D) = \mathrm{Tr}(DX_R) - \mathrm{Tr}(DX_D) \geqslant 0,$$

where the last step follows from the optimality of X_D. Similarly, $\mathrm{Tr}(\widetilde{R}X_D) - \mathrm{Tr}(\widetilde{R}X_R) = \mathrm{Tr}(R(X_D - X_R))$, and so (6.8) implies the result. □

Now it suffices to bound $\mathcal{F}(\widetilde{D}-\widetilde{R})$. For an $n_1 \times n_2$ matrix X, consider the matrix norm

$$\|X\|_{1,\infty} := \sum_{i=1}^{n_1} \max_{1\leqslant j\leqslant n_2} |X_{i,j}| = \sum_{i=1}^{n_1} \|X_{i,.}\|_\infty.$$

The following lemma will be useful:

Lemma 6.3.3. $\mathcal{F}(M) \leqslant \min\{\|M\|_{1,\infty}, \min\{k,r\}\|M\|_{2\to 2}\}$ *where* $r = \mathit{rank}(M)$.

Proof. The first bound follows from the classical version of Hölder's inequality (recalling that $X_{i,j} \geqslant 0$ and $X1 = 1$ by design):

$$|\mathrm{Tr}(MX)| \leqslant \sum_{i=1}^{N}\sum_{j=1}^{N} |M_{i,j}X_{i,j}| \leqslant \sum_{i=1}^{N} \|M_{i,\cdot}\|_\infty \left(\sum_{j=1}^{N} |X_{i,j}|\right) = \sum_{i=1}^{N} \|M_{i,\cdot}\|_\infty$$

The second bound is a consequence of Von Neumann's trace inequality: if the singular values of X and M are respectively $\alpha_1 \geqslant \ldots \geqslant \alpha_N$ and $\beta_1 \geqslant \ldots \geqslant \beta_N$ then

$$|\mathrm{Tr}(MX)| \leqslant \sum_{i=1}^{N} \alpha_i \beta_i$$

Since X is feasible in (6.7) we have $\alpha_1 \leqslant 1$ and $\sum_{i=1}^{N} \alpha_i \leqslant k$. Using that $\mathrm{rank}(M) = r$ we get

$$|\mathrm{Tr}(MX)| \leqslant \sum_{i=1}^{k} \beta_i \leqslant \min\{k, r\}\|M\|_{2\to 2} \qquad \square$$

Proof of Theorem 6.2.2. Write $x_{t,i} = r_{t,i} + \gamma_t$. Then

$$\begin{aligned}(D_{ab})_{ij} &= \|x_{a,i} - x_{b,j}\|_2^2 \\ &= \|(r_{a,i} + \gamma_a) - (r_{b,j} + \gamma_b)\|_2^2 \\ &= \|r_{a,i} - r_{b,j}\|_2^2 + 2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle + \|\gamma_a - \gamma_b\|_2^2.\end{aligned}$$

Furthermore,

$$\|r_{a,i} - r_{b,j}\|_2^2 = \|r_{a,i}\|_2^2 - 2\langle r_{a,i}, r_{b,j}\rangle + \|r_{b,j}\|_2^2 = ((\mu 1^\top + G^\top G + 1\mu^\top)_{ab})_{ij},$$

where G is the matrix whose (a,i)th column is $r_{a,i}$, and μ is the column vector whose (a,i)th entry is $\|r_{a,i}\|_2^2$. Recall that

$$(R_{ab})_{ij} = \xi + \Delta_{ab}^2/2 + \max\left\{0, \Delta_{ab}^2/2 + 2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle\right\}.$$

Then $P_{1^\perp}(D-R)P_{1^\perp} = P_{1^\perp}G^\top G P_{1^\perp} + P_{1^\perp}FP_{1^\perp}$ where

$$(F_{ab})_{ij} = \begin{cases} \Delta_{ab}^2/2 + 2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle & \text{if } 2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle \leqslant -\Delta_{ab}^2/2 \\ 0 & \text{otherwise.} \end{cases}$$

Considering Lemma 6.3.3 and that $\operatorname{rank}(G^\top G) \leqslant m$ we will bound

$$\mathcal{F}(M) \leqslant \min\{k, m\}\|P_{1^\perp}G^\top G P_{1^\perp}\|_{2\to 2} + \frac{1}{n_{\min}}\|P_{1^\perp}FP_{1^\perp}\|_{1,\infty}. \tag{6.9}$$

For the first term we observe $\|P_{1^\perp}G^\top G P_{1^\perp}\|_{2\to 2} \leqslant \|G^\top G\|_{2\to 2} = \|G^\top\|_{2\to 2}^2$. Note that if the rows $X_i^{(t)}$, $i = 1, \dots n_t$ of $G^\top$ come from a distribution with second moment matrix Σ_t, then $X_i^{(t)}$ has the same distribution as $\Sigma_t^{1/2} g$, where g is an isotropic random vector. Then $\|G^\top\| \leqslant \sigma_{\max}\|\tilde{G}^\top\|$ where the rows of $\tilde{G}^\top$ are isotropic random vectors.

By Theorem 5.39 in [64], we have that there exist c_1 and c_2 constants depending only on the subgaussian norm of the rows of G such that with probability $\geqslant 1-\eta$:

$$\|G^\top\|_{2\to 2} \leqslant \sigma_{\max}\left(\sqrt{N} + c_1\sqrt{m} + \sqrt{c_2 \log(2/\eta)}\right).$$

Note that by Corollary 3.35, when the rows of $G^\top$ are Gaussian random vectors we have the result for $c_1 = 1$ and $c_2 = 2$.

For bounding the second term in (6.9), the triangle inequality gives $\|P_{1^\perp}FP_{1^\perp}\|_{1,\infty} \leqslant 4\|F\|_{1,\infty}$. In order to get a handle on $\|F\|_{1,\infty}$ we first compute

the expected value of its entries using that $|2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle|$ obeys a folded subgaussian distribution, coming from a subgaussian with variance at most $8\sigma_{\max}^2\Delta_{ab}^2$:

$$\begin{aligned}
&\mathbb{E}|(F_{ab})_{ij}| \\
&\quad\leqslant \left(\Delta_{ab}^2/2 + \mathbb{E}|2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle|\right) \mathbb{P}\left(2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle < -\Delta_{ab}^2/2\right) \\
&\quad\leqslant \left(\frac{\Delta_{ab}^2}{2} + \frac{4\sigma_{\max}\Delta_{ab}}{\sqrt{\pi}}\right) \exp\left(-\frac{\Delta_{ab}^2}{64\sigma_{\max}^2}\right) \\
&\quad\leqslant \Delta_{ab}^2 \exp\left(-\frac{\Delta_{ab}^2}{64\sigma_{\max}^2}\right) \text{ assuming } \Delta_{\min}^2 > 16k\sigma_{\max}^2, \quad k \geqslant 2 \\
&\quad\leqslant \Delta_{ab}^2 \frac{64^2\sigma_{\max}^4}{\Delta_{ab}^4} \text{ using } e^{-x} \leqslant \frac{1}{x^2} \text{ for } x > 0. \\
&\quad\leqslant -\frac{256\sigma_{\max}^2}{k} \text{ using again } \Delta_{\min}^2 > 16k\sigma_{\max}^2, \quad k \geqslant 2 \\
&\quad= O(\sigma_{\max}^2/k)
\end{aligned}$$

Now we can write $F = 2(L - L^\top)$ where $L_{a,i} := (L_{ab})_{ij} \in \{\langle r_{a,i}, \gamma_a - \gamma_b\rangle, 0\}$ has independent rows, and $\mathbb{E}|(L_{ab})_{ij}| \leqslant \mathbb{E}|(F_{ab})_{ij}| = O(\sigma_{\max}^2/k)$. We can then bound

$$\|F\|_{1,\infty} \leqslant 4\|L\|_{1,\infty} \leqslant \|L^{small}\|_{1,1}$$

where $L^{small} \in \mathbb{R}^{N\times k}$ is a submatrix of distinct columns.

Then we have a high-probability estimate:

$$\begin{aligned}
\mathbb{P}(\|L^{small}\|_{1,1} > t) &\leqslant \mathbb{P}\left(2k\sum_{a=1}^{k}\sum_{i=1}^{n_a}|L_{a,i}| > t\right) \\
&\leqslant \mathbb{P}\left(\sum_{a=1}^{k}\sum_{i=1}^{n_a}(|L_{a,i}| - \mathbb{E}|L_{a,i}|) > \frac{t}{2k} - c_3\sigma_{\max}^2 n_{\max}\right).
\end{aligned}$$

Using that $L_{a,i}$ are independent subgaussian random variables, we know there exist constants $c_4, c_5 \geqslant 0$ such that

$$\mathbb{P}\left(\sum_{a=1}^{k}\sum_{i=1}^{n_a}(|L_{a,i}| - \mathbb{E}|L_{a,i}|) > u\right) \leqslant c_4 \exp\left(-c_5\frac{u^2}{N}\right).$$

So, choosing $t = 2c_3 k n_{\max}\sigma_{\max}^2 + \sqrt{\frac{N}{c_5}\log\frac{c_4}{\eta}}$, we get that with probability at least $1-\eta$

$$\|P_{1^\perp} F P_{1^\perp}\|_{1,\infty} \leqslant 8c_3 k n_{\max}\sigma_{\max}^2 + 4\sqrt{\frac{N}{c_5}\log\frac{c_4}{\eta}}$$

Putting everything together, we get that there exist constants C_1, C_2, C_3 such that with probability at least $1-2\eta$

$$\begin{aligned}
\|X_D - X_R\|_F^2 &\leqslant \frac{5}{n_{\min}\Delta_{\min}^2}\operatorname{Tr}(R(X_D - X_R)) \\
&\leqslant \frac{10}{n_{\min}\Delta_{\min}^2}\mathcal{F}(\widetilde{D} - \widetilde{R}) \\
&\leqslant C_1\frac{\min\{k,m\}\left(\sqrt{N} + c_1\sqrt{m} + \sqrt{c_2\log(2/\eta)}\right)^2\sigma_{\max}^2}{n_{\min}\Delta_{\min}^2} \\
&\qquad + C_2\frac{k n_{\max}\sigma_{\max}^2}{n_{\min}\Delta_{\min}^2} + C_3\frac{\sqrt{N\log c_4/\eta}}{n_{\min}\Delta_{\min}^2}.
\end{aligned}$$

If additionally we require $N > \max\{c_1 m, c_2\log(2/\eta), \log(c_4/\eta)\}$, we get

$$\|X_D - X_R\|_F^2 \leqslant C\frac{k\alpha\sigma_{\max}^2(\alpha + \min\{k,m\})}{\Delta_{\min}^2}.$$

Rearranging gives the result. □

6.4 Denoising

In the special case where each Gaussian is spherical with the same entrywise variance σ^2 and the same number n of samples, Theorem 6.2.2 says:

$$\|X_D - X_R\|_F^2 \lesssim \frac{k^2\sigma^2}{\Delta_{\min}^2}$$

with high probability as $n \to \infty$.

Let P denote the $m \times N$ matrix whose (a,i)th column is $x_{a,i}$. Then PX_R is an $m \times N$ matrix whose (a,i)th column is $\tilde{\gamma}_a$, a good estimate of γ_a, and so one might expect PX_D to have its columns be close to the $\tilde{\gamma}_a$'s. This is precisely what the following theorem gives:

Theorem 6.4.1. *Suppose $\sigma \lesssim \Delta_{\min}/\sqrt{k}$. Let P denote the $m \times N$ matrix whose (a,i)th column is $x_{a,i}$, and let $c_{a,i}$ denote the (a,i)th column of PX_D. Then*

$$\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n}\|c_{a,i} - \tilde{\gamma}_a\|_2^2 \lesssim \frac{\|\Gamma\|_{2\to 2}^2}{\Delta_{\min}^2} \cdot k\sigma^2$$

with high probability as $n \to \infty$. Here, the ath column of Γ is $\tilde{\gamma}_a - \frac{1}{k}\sum_{b=1}^{k}\tilde{\gamma}_b$.

The proof can be found at the end of this section. For comparison,

$$\mathbb{E}\left[\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n}\|x_{a,i} - \gamma_a\|_2^2\right] = m\sigma^2, \tag{6.10}$$

meaning the $c_{a,i}$'s serve as "denoised" versions of the $x_{a,i}$'s provided $\|\Gamma\|_{2\to 2}$ is not too large compared to $\Delta_{\min}$. The following lemma investigates this provision:

Lemma 6.4.2. *For every choice of* $\{\tilde{\gamma}_a\}_{a=1}^k$*, we have*

$$\frac{\|\Gamma\|_{2\to 2}^2}{\Delta_{\min}^2} \geqslant \frac{1}{2},$$

with equality if $\{\tilde{\gamma}_a\}_{a=1}^k$ *is a simplex. More generally, if the following are satisfied simultaneously:*

(i) $\sum_{a=1}^k \tilde{\gamma}_a = 0$*,*

(ii) $\|\tilde{\gamma}_a\|_2 \asymp 1$ *for every* $a \in \{1,\dots,k\}$*, and*

(iii) $|\langle \tilde{\gamma}_a, \tilde{\gamma}_b \rangle| \lesssim 1/k$ *for every* $a, b \in \{1,\dots,k\}$ *with* $a \neq b$*,*

then

$$\frac{\|\Gamma\|_{2\to 2}^2}{\Delta_{\min}^2} \lesssim 1.$$

See the end of the section for the proof. Plugging these estimates for $\|\Gamma\|_{2\to 2}^2/\Delta_{\min}^2$ into Theorem 6.4.1 shows that the $c_{a,i}$'s in this case exhibit denoising to an extent that the m in (6.10) can be replaced with k:

$$\frac{1}{N}\sum_{a=1}^k \sum_{i=1}^n \|c_{a,i} - \tilde{\gamma}_a\|_2^2 \lesssim k\sigma^2.$$

For more general choices of $\{\tilde{\gamma}_a\}_{a=1}^k$, one may attempt to estimate $\|\Gamma\|_{2\to 2}$ in terms of $\Delta_{\max}$, but this comes with a bit of loss in the denoising estimate:

Corollary 6.4.3. *If* $k\sigma \lesssim \Delta_{\min} \leqslant \Delta_{\max} \lesssim K\sigma$*, then*

$$\frac{1}{N}\sum_{a=1}^k \sum_{i=1}^n \|c_{a,i} - \tilde{\gamma}_a\|_2^2 \lesssim K^2\sigma^2$$

with high probability as $n \to \infty$*.*

Indeed, this doesn't guarantee denoising unless $k \lesssim K \leqslant \sqrt{m}$. To prove this corollary, apply the following string of inequalities to Theorem 6.4.1:

$$\|\Gamma\|_{2\to 2}^2 \leqslant \|\Gamma\|_F^2 \leqslant k\Delta_{\max}^2 \lesssim kK^2\sigma^2,$$

where the second inequality uses the following lemma:

Lemma 6.4.4. *If* $\sum_{a=1}^{k} \tilde{\gamma}_a = 0$, *then* $\|\tilde{\gamma}_a\|_2 \leqslant \Delta_{\max}$ *for every* a.

Proof. Fix a. Then

$$\begin{aligned}\min_{b\neq a} \left\langle \tilde{\gamma}_b, \frac{\tilde{\gamma}_a}{\|\tilde{\gamma}_a\|_2} \right\rangle &\leqslant \frac{1}{k-1} \sum_{\substack{b=1 \\ b\neq a}}^{k} \left\langle \tilde{\gamma}_b, \frac{\tilde{\gamma}_a}{\|\tilde{\gamma}_a\|_2} \right\rangle \\ &= \frac{1}{k-1} \left\langle \sum_{b=1}^{k} \tilde{\gamma}_b - \tilde{\gamma}_a, \frac{\tilde{\gamma}_a}{\|\tilde{\gamma}_a\|_2} \right\rangle = -\frac{1}{k-1} \|\tilde{\gamma}_a\|_2.\end{aligned}$$

Let $b(a)$ denote the minimizer. Then Cauchy–Schwarz gives

$$\begin{aligned}\Delta_{\max} &\geqslant \|\tilde{\gamma}_a - \tilde{\gamma}_{b(a)}\|_2 \geqslant \left\langle \tilde{\gamma}_a - \tilde{\gamma}_{b(a)}, \frac{\tilde{\gamma}_a}{\|\tilde{\gamma}_a\|_2} \right\rangle \\ &\geqslant \|\tilde{\gamma}_a\|_2 + \tfrac{1}{k-1}\|\tilde{\gamma}_a\|_2 \geqslant \|\tilde{\gamma}_a\|_2.\end{aligned}$$

□

Proof of Theorem 6.4.1. Without loss of generality, we have $\sum_{a=1}^{k} \tilde{\gamma}_a = 0$. Write

$$\sum_{a=1}^{k} \sum_{i=1}^{n} \|c_{a,i} - \tilde{\gamma}_a\|_2^2 = \|P(X_D - X_R)\|_F^2 \leqslant \|P\|_{2\to 2}^2 \|X_D - X_R\|_F^2. \tag{6.11}$$

Decompose $P = \Gamma \otimes 1^\top + G$, where 1 is n-dimensional and G has i.i.d. entries from $\mathcal{N}(0, \sigma^2)$. Observe that

$$\|\Gamma \otimes 1^\top\|_{2\to 2}^2 = \|(\Gamma \otimes 1^\top)(\Gamma \otimes 1^\top)^\top\|_{2\to 2} = \|n\Gamma\Gamma^\top\|_{2\to 2} = n\|\Gamma\|_{2\to 2}^2. \tag{6.12}$$

Also, Corollary 5.35 in [64] gives that

$$\|G\|_{2\to 2} \lesssim (\sqrt{N} + \sqrt{m})\sigma \lesssim \sqrt{N}\sigma \tag{6.13}$$

with probability $\geqslant 1 - e^{-\Omega_m(N)}$. The result then follows from estimating $\|P\|_{2\to 2}$ with (6.12) and (6.13) by triangle inequality, plugging into (6.11), and then applying Theorem 6.2.2. □

Proof of Lemma 6.4.2. Since $\|\Gamma x\|_2 \leqslant \|\Gamma\|_{2\to 2}\|x\|_2$ for every x, we have that

$$\|\Gamma\|_{2\to 2}^2 \geqslant \frac{\|\tilde{\gamma}_a - \tilde{\gamma}_b\|_2^2}{2}$$

for every a and b, and so

$$\frac{\|\Gamma\|_{2\to 2}^2}{\Delta_{\min}^2} \geqslant \frac{1}{2} \cdot \frac{\Delta_{\max}^2}{\Delta_{\min}^2} \geqslant \frac{1}{2}.$$

For the second part, let $\{\tilde{\gamma}_a\}_{a=1}^k$ be a simplex. Without loss of generality, $\{\tilde{\gamma}_a\}_{a=1}^k$ is centered at the origin, each point having unit 2-norm. Then $\langle \tilde{\gamma}_1, \tilde{\gamma}_2 \rangle = -1/(k-1)$, and so

$$\Delta_{\min}^2 = \|\tilde{\gamma}_1 - \tilde{\gamma}_2\|_2^2 = \|\tilde{\gamma}_1\|_2^2 + \|\tilde{\gamma}_2\|_2^2 - 2\langle \tilde{\gamma}_1, \tilde{\gamma}_2 \rangle = \frac{2k}{k-1}.$$

Next, we write

$$\Gamma^\top \Gamma = \frac{k}{k-1} I - \frac{1}{k-1} 11^\top,$$

and conclude that $\|\Gamma\|_{2\to 2}^2 = \|\Gamma^\top \Gamma\|_{2\to 2} = k/(k-1)$. Combining with our expression for $\Delta_{\min}^2$ then gives the result. For the last part, pick a and b such that $\Delta_{\min} = \|\tilde{\gamma}_a - \tilde{\gamma}_b\|_2$. Then

$$\Delta_{\min}^2 = \|\tilde{\gamma}_a\|_2^2 + \|\tilde{\gamma}_b\|_2^2 - 2\langle \tilde{\gamma}_a, \tilde{\gamma}_b \rangle \gtrsim 2 - 2/k.$$

Also, Gershgorin implies

$$\|\Gamma\|_{2\to 2}^2 = \|\Gamma^\top \Gamma\|_{2\to 2} \lesssim 1 + (k-1)/k,$$

and so combining these estimates gives the result. □

6.5 Rounding

Theorem 6.5.1. *Take $\epsilon < \Delta_{\min}/8$, suppose*

$$\#\Big\{(a,i) : \|c_{a,i} - \tilde{\gamma}_a\|_2 > \epsilon\Big\} < \frac{n}{2},$$

and consider the graph G of vertices $\{c_{a,i}\}_{i=1,\,a=1}^{n\ \ k}$ such that $c_{a,i} \leftrightarrow c_{b,j}$ if $\|c_{a,i} - c_{b,j}\|_2 \leqslant 2\epsilon$. For each $i = 1,\dots,k$, select the vertex v_i of maximum degree (breaking ties arbitrarily) and update G by removing every vertex w such that $\|w - v_i\|_2 \leqslant 4\epsilon$. Then there exists a permutation π on $\{1,\dots,k\}$ such that

$$\|v_i - \tilde{\gamma}_{\pi(i)}\|_2 \leqslant 3\epsilon$$

for every $i \in \{1,\dots,k\}$.

Proof. Let $B(x,r)$ denote the closed 2-ball of radius r centered at x. For each i, we will determine $\pi(i)$ at the conclusion of iteration i. Denote $R_1 := \{1,\dots,k\}$ and $R_{i+1} := R_i \setminus \{\pi(i)\}$ for each $i = 2,\dots,k-1$. We claim that the following hold at the beginning of each iteration i:

(i) $< n/2$ vertices lie outside $\bigcup_{a\in R_i} B(\tilde{\gamma}_a, \epsilon)$,

(ii) $\geqslant n/2$ vertices lie inside $B(v_i, 2\epsilon)$, and

(iii) there exists a unique $a \in R_i$ such that $\|v_i - \tilde{\gamma}_a\|_2 \leqslant 3\epsilon$.

First, we show that for each i, (i) and (ii) together imply (iii). Indeed, there are enough vertices in $B(v_i, 2\epsilon)$ that one of them must reside in $B(\tilde{\gamma}_a, \epsilon)$ for some $a \in R_i$. Furthermore, this a is unique since $\epsilon < \Delta_{\min}/6$. By triangle inequality, we have $\|v_i - \tilde{\gamma}_a\|_2 \leqslant 3\epsilon$, and so we put $\pi(i) := a$.

We now prove (i) and (ii) by induction. When $i = 1$, we have (i) by assumption. For (ii), note that each $B(\tilde{\gamma}_a, \epsilon)$ contains $\geqslant n/2$ of the vertices, and by triangle inequality, each has degree $\geqslant n/2 - 1$ in G. As such, the vertex v_1 of maximum degree will have degree $\geqslant n/2 - 1$, thereby implying (ii).

Now suppose (i), (ii) and (iii) all hold for iteration $i < k$. By triangle inequality, (iii) implies $B(\tilde{\gamma}_{\pi(i)}, \epsilon) \subseteq B(v_i, 4\epsilon)$. As such, the ith iteration removes all vertices in $B(\tilde{\gamma}_{\pi(i)}, \epsilon)$ so that (i) continues to hold for iteration $i + 1$. Next, $\epsilon < \Delta_{\min}/8$ and (iii) together imply that the removal of vertices in $B(v_i, 4\epsilon)$ preserves the vertices in $B(\tilde{\gamma}_a, \epsilon)$ for every $a \in R_{i+1}$, and their degrees are still $\geqslant n/2 - 1$ by the same triangle argument as before. Thus, (ii) holds for iteration $i + 1$. □

Corollary 6.5.2. *Suppose* $k \lesssim m$, *and denote* $S := \|\Gamma\|_{2\to 2}/\Delta_{\min}$. *Pick* $\epsilon \asymp Sk\sigma$. *Perform the rounding scheme of Theorem 6.5.1 to columns of* PX_D. *Then with high probability,* $\{v_i\}_{i=1}^k$ *satisfies*

$$\|v_i - \tilde{\gamma}_{\pi(i)}\|_2 \lesssim Sk\sigma$$

for some permutation π, provided $\sigma \lesssim \Delta_{\min}/(Sk)$.

By Lemma 6.4.2, we have $S \lesssim 1$ in the best-case scenario. In this case, our rounding scheme works in the regime $\sigma \lesssim \Delta_{\min}/k$. (Note that denoising is guaranteed in the regime $\sigma \lesssim \Delta_{\min}/\sqrt{k}$). In general, the cost of rounding is a factor of k in the average squared deviation of our estimates:

$$\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n}\|c_{a,i}-\tilde{\gamma}_a\|_2^2 \lesssim S^2k\sigma^2, \quad \text{whereas} \quad \frac{1}{k}\sum_{i=1}^{k}\|\nu_i-\tilde{\gamma}_{\pi(i)}\|_2^2 \lesssim S^2k^2\sigma^2.$$

On the other hand, we are not told which of the points in $\{c_{a,i}\}_{i=1,\,a=1}^{n\quad k}$ correspond to any given $\tilde{\gamma}_a$, whereas in rounding, we know that each ν_i corresponds to a distinct $\tilde{\gamma}_a$.

Proof of Corollary 6.5.2. Draw (a,i) uniformly from $\{1,\dots,k\}\times\{1,\dots,n\}$ and take X to be the random variable $\|c_{a,i}-\tilde{\gamma}_a\|_2^2$. Then Markov's inequality and Theorem 6.4.1 together give

$$\begin{aligned}\#\Big\{(a,i):\|c_{a,i}-\tilde{\gamma}_a\|_2>\epsilon\Big\} &= N\cdot\mathbb{P}(X>\epsilon^2)\\ &\leqslant \frac{N}{\epsilon^2}\cdot\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n}\|c_{a,i}-\tilde{\gamma}_a\|_2^2 \lesssim \frac{N}{\epsilon^2}\cdot S^2k\sigma^2 \lesssim \frac{n}{2}.\end{aligned}$$

For Theorem 6.5.1 to apply, it suffices to ensure $\epsilon < \Delta_{\min}/8$, which follows from $\sigma \lesssim \Delta_{\min}/(Sk)$. □

6.5.1 Numerical example: Clustering the MNIST dataset

In this section, we apply our clustering algorithm to the NMIST handwritten digits dataset [40]. This dataset consists of 70,000 different 28×28

grayscale images, reshaped as 784×1 vectors; 55,000 of them are considered training set, 10,000 are test set, and the remaining 5,000 are validation set.

Clustering the raw data gives poor results (due to 4's and 9's being similar, for example), so we first learn meaningful features, and then cluster the data in feature space. To simplify feature extraction, we used the first example from the TensorFlow tutorial [1]. This consists of a one-layer neural network $y(x) = \sigma(Wx + b)$, where σ is the softmax function, W is a 784×10 matrix to learn, and b is a 10×1 vector to learn. As the tutorial shows, the neural network is trained for 1,000 iterations, each iteration using batches of 100 random points from the training set.

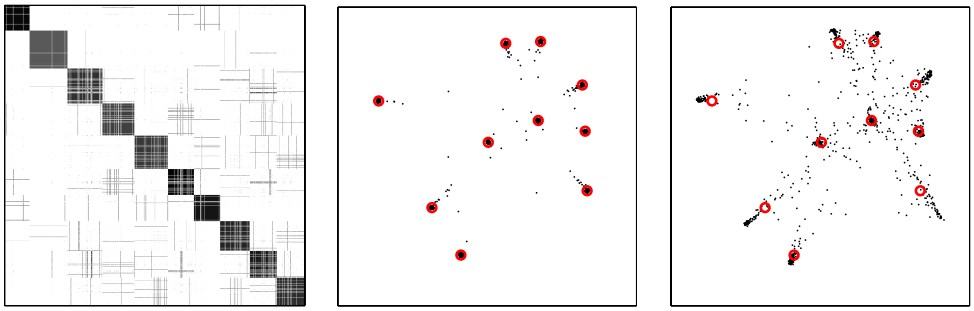
Training the neural network amounts to finding W and b that fit the training set well. After selecting these parameters, we run the trained neural network on the first 1,000 elements of the test set, obtaining $\{y(x_i)\}_{i=1}^{1000}$, where each $y(x_i)$ is a 10×1 vector representing the probabilities of being each digit. Since $y(x_i)$ is a probability vector, its entries sum to 1, and so the feature space is actually 9-dimensional.

For this experiment, we cluster $\{y(x_i)\}_{i=1}^{1000}$ with two different algorithms: (i) MATLAB's built-in implementation of k-means++, and (ii) our relax-and-round algorithm based on the k-means semidefinite relaxation (k-means sdp). (The results of the latter alternative are illustrated in Figure 6.1.)

Since each run of k-means++ uses a random initialization that im-

pacts the partition, we ran this algorithm 100 times. In fact, the k-means value of the output varied quite a bit: the all-time low was 39.1371, the all-time high was 280.4174, and the median was 108.2358; the all-time low was reached in 34 out of the 100 trials. Since our relax-and-round alternative has no randomness, the outcome is deterministic, and its k-means value was 39.1371, i.e., identical to the all-time low from k-means++. By comparison, the k-means value of the planted solution (i.e., clustering according to the hidden digit label) was 103.5430, and the value of the SDP (which serves as a lower bound on the optimal k-means value) was 38.5891. As such, not only did our relax-and-round alternative produce the best clustering that k-means++ could find, it also provided a certificate that no clustering has a k-means value that is 1.5% better.

Recalling the nature of our approximation guarantees, we also want to know well the relax-and-round algorithm's clustering captures the ground truth. To evaluate this, we determined a labeling of the clusters for which the resulting classification exhibited a minimal misclassification rate. (This amounts to minimizing a linear objective over all permutation matrices, which can be relaxed to a generically tight linear program over doubly stochastic matrices.) For k-means++, the all-time low misclassification rate was 0.0971 (again, accomplished by 34 of the 100 trials), the all-time high was 0.4070, and the median was 0.2083. As one might expect, the relax-and-round output had a misclassification rate of 0.0971.



(a) (b) (c)

Figure 6.1: Clustering MNIST with k-means SDP.
(a) After applying TensorFlow [1] to learn a 9-dimensional feature space of MNIST digits [40], determine the features of the first 1,000 images in the MNIST test set, compute the 1000×1000 matrix D of squared distances in feature space, and then solve the k-means semidefinite relaxation (k-means sdp) using SDPNAL+v0.3 [71]. (The computation takes about 6 minutes on a standard MacBook Air laptop.) Convert the SDP-optimizer X to a grayscale image such that white pixels denote zero entries. By inspection, this matrix is not exactly of the form (1.8), but it looks close, and it certainly appears to have low rank. **(b)** Letting P denote the 9×1000 matrix whose columns are the feature vectors to cluster, compute the denoised data PX and identify the 10 most popular locations in $\mathbb{R}^9$ (denoted by red circles) among the columns of PX (denoted by black dots). For the plot, we project the 9-dimensional data onto a random 2-dimensional subspace. **(c)** The 10 most popular locations form our estimates for the centers of digits in feature space. We plot these locations relative to the original data, projected in the same 2-dimensional subspace as (b).

Chapter 7

Polynomial-time lower bound of NP-hard functions

We begin by recalling the definition of the Gromov-Hausdorff distance introduced in Section 1.2; for this, we start with the Hausdorff distance. Let (Z, d) a compact metric space and $\mathcal{C}(Z)$ the collection of all compact sets in Z. If $A, B \in \mathcal{C}(Z)$, the Hausdorff distance between A and B can be expressed as

$$d_{\mathcal{H}}^{Z}(A, B) = \inf_{R \in \mathcal{R}(A,B)} \sup_{(a,b)\in R} d(a, b)$$

where $\mathcal{R}(A, B)$ is the set of correspondences in $R \subset A \times B$ such that every element $a \in A$ is related to at least one element in B and every element $b \in B$ is related at least one element in A. For many theoretical and practical applications it is common to relax such distance to the Wasserstein distance [65]. In that setting, one endows $\mathcal{C}(Z)$ with a measure,

$$\mathcal{C}_w(Z) = \{(A, \mu_A) \colon A \in \mathcal{C}(Z),\ \mathrm{supp}(\mu_A) = A\},$$

This chapter is based on the article:
Soledad Villar, Afonso S. Bandeira, Andrew Blumberg, Rachel Ward. *A polynomial-time relaxation of the Gromov-Hausdorff distance*, 2016 (submitted).
The author was the main contributor to the entire article.

and relaxes the set of correspondences $\mathcal{R}(A, B)$ to the set of transportation plans

$$\mathcal{M}(\mu_A, \mu_B) = \{\mu: \quad \mu(A_0 \times B) = \mu_A(A_0), \quad \mu(A \times B_0) = \mu_B(B_0),$$
$$\text{for all Borel sets } A_0 \subset A,\ B_0 \subset B\}.$$

Then, for $A, B \in \mathcal{C}_w(Z)$, the Wasserstein distance is defined for $1 \leqslant p \leqslant \infty$ as

$$d^Z_{W,p}(A, B) = \inf_{\mu \in \mathcal{M}(\mu_A, \mu_B)} \left(\int_{A \times B} d^p(a, b) \, d\mu(a, b) \right)^{1/p} \quad \text{for } 1 \leqslant p < \infty$$

$$d^Z_{W,\infty}(A, B) = \inf_{\mu \in \mathcal{M}(\mu_A, \mu_B)} \sup_{(a,b) \in \text{supp}(\mu)} d(a, b).$$

For A, B finite sets this distance can be efficiently computed by a linear program.

The Hausdorff distance suffices to compare metric spaces embedded in a common ambient metric space; Gromov's idea to extend this to compare arbitrary metric spaces is simply to consider the infimum over all isometric embeddings into a common metric space [31]. Specifically, if X, Y are compact metric spaces, the Gromov-Hausdorff distance is defined as

$$d_{\mathcal{GH}}(X, Y) = \inf_{Z, f, g} d^Z_{\mathcal{H}}(f(X), g(Y))$$

where $f\colon X \to Z$ and $g\colon Y \to Z$ are isometric embeddings into Z, a metric space. Unfortunately, it is an NP-hard problem to compute this distance.

Since the Hausdorff distance becomes computationally tractable when relaxed to the Wasserstein distance, one might consider a transport-based relaxation of the Gromov-Hausdorff distance that works in the setting of

metric measure spaces. In a series of articles [47, 46, 48] Mémoli considers different equivalent expressions for the Gromov-Hausdorff distance, and by relaxing them and considering them in the measure metric space setting, he obtains different *gromovizations* of the Wasserstein distance, called Gromov-Wasserstein distances. A particularly natural relaxation is based on the observation that the Gromov-Hausdorff distance can be expressed as:

$$d_{\mathcal{GH}}(X,Y) = \frac{1}{2} \inf_{R \in \mathcal{R}(X,Y)} \sup_{\substack{x,x' \in X \\ y,y' \in Y \\ (x,y),(x',y') \in R}} \Gamma_{X,Y}(x,y,x',y') \tag{7.1}$$

where $\Gamma_{X,Y}(x,y,x',y') = |d_X(x,x') - d_Y(y,y')|$. For $1 \leqslant p \leqslant \infty$ Mémoli then defines Gromov-Wasserstein relaxations of the Gromov-Hausdorff distance as

$$D_p(X,Y) = \frac{1}{2} \inf_{\mu \in \mathcal{M}(\mu_X,\mu_Y)} \left(\int_{X \times Y} \int_{X \times Y} \left(\Gamma_{X,Y}(x,y,x',y') \right)^p \mu(dx \times dy) \mu(dx' \times dy') \right)^{1/p} \tag{7.2}$$

$$D_\infty(X,Y) = \frac{1}{2} \inf_{\mu \in \mathcal{M}(\mu_X,\mu_Y)} \sup_{\substack{x,x' \in X \\ y,y' \in Y \\ (x,y),(x',y') \in \mathrm{supp}(\mu)}} \Gamma_{X,Y}(x,y,x',y') \tag{7.3}$$

In his work, Mémoli studies topological properties of the different distance relaxations and how they compare with each other and with the Gromov-Hausdorff distance. He also proposes an algorithm to approximate D_p, but due the non-convexity of its objective function, no performance guarantees are provided. Recent work [57] has provided efficient heuristic algorithms based on optimal transportation to approximate the

Gromov-Wasserstein distance for alignment applications. An efficient spectral interpretation of the Gromov-Hausdorff distance has been recently proposed for matching surfaces [5] as well.

Remark 7.0.1. Gromov considered another metric on the set of metric measure spaces defined in terms of the convergence of all distance matrix distributions (i.e., the distributions induced by taking the pushforward of the measure to a collection of n points and applying the metric to all pairs). It turns out that this metric is closely related to the Gromov-Wasserstein distance [59, 3.7]. Moreover, these metrics induce the same notion of convergence as arises in the theory of dense graph sequences and graphons. Specifically, we can regard a graph as a metric measure space; the underlying metric space has points the set of vertices and pairwise distances $\frac{1}{2}$ if the points are connected and 1 otherwise, and the measure assigns equal mass to each point. (See [26] for further discussion of this point).

7.1 Semidefinite programming relaxations of Gromov-Wasserstein and Gromov-Hausdorff distances

Consider the setting where X and Y are finite metric spaces (or metric measure spaces), say $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ (with measures $\mu_X(x_i) = v_i$ and $\mu_Y(y_j) = \lambda_j$). Let us abbreviate $\Gamma_{X,Y}(x_i, y_j, x_{i'}, y_{j'})$ as $\Gamma_{ij,i'j'}$ for $i, i' = 1, \ldots, n$ and $j, j' = 1, \ldots, m$. The formulation of the Gromov-Hausdorff distance given in equation (7.1) can be expressed as a quadratic assignment

(Remark 3 in [46]):

$$d_{\mathcal{GH}}(X,Y) = \frac{1}{2}\min_{\mu}\max_{ij,i'j'}\Gamma_{ij,i'j'}\mu_{ij}\mu_{i'j'} \tag{7.4}$$
$$\text{subject to } \mu_{ij}\in\{0,1\},\ \sum_{j=1}^{m}\mu_{ij}\geqslant 1,\ \sum_{i=1}^{n}\mu_{ij}\geqslant 1$$

and the expressions (7.2) and (7.3) can be written as (7.5) and (7.6) respectively:

$$D_p(X,Y) = \frac{1}{2}\min_{\mu\in\mathbb{R}^{n\times m}}\left(\sum_{i,i'=1}^{n}\sum_{j,j'=1}^{m}\mu_{ij}\mu_{i'j'}\Gamma^p_{ij,i'j'}\right)^{1/p} \tag{7.5}$$
$$\text{subject to } 0\leqslant\mu_{ij}\leqslant 1, \sum_{i=1}^{n}\mu_{ij}=\lambda_j, \sum_{j=1}^{m}\mu_{ij}=\nu_i$$

$$D_\infty(X,Y) = \frac{1}{2}\min_{\mu\in\mathbb{R}^{n\times m}}\max_{\substack{i,i',j,j'\\ \mu_{ij}\mu_{i'j'}\neq 0}}\Gamma_{ij,i'j'} \tag{7.6}$$
$$\text{subject to } 0\leqslant\mu_{ij}\leqslant 1, \sum_{i=1}^{n}\mu_{ij}=\lambda_j, \sum_{j=1}^{m}\mu_{ij}=\nu_i$$

In order to approach non-convex optimization problems like (7.4), (7.5) or (7.6), one standard technique is to linearize the objective by lifting $\mu_{ij}\mu_{i'j'}$ and μ_{ij} to a symmetric variable $\mathbf{Z}\in\mathbb{R}^{nm+1\times nm+1}$ whose entries entries are indexed by pairs $(ij,i'j')$,$(ij,nm+1)$, $(nm+1,i'j')$ and $(nm+1,nm+1)$ with $i,i'=1,\dots n$ and $j,j'=1,\dots,m$.

$$\mathbf{Z} = \begin{bmatrix} \hat{\mathbf{Z}} & \mathbf{z} \\ \mathbf{z}^\top & 1 \end{bmatrix}. \tag{7.7}$$

Then note that, for instance, the problems (7.5) and (7.6) are equivalent to problems (7.8) and (7.9) respectively:

$$D_p(X,Y) = \frac{1}{2}\left(\min_{\mathbf{Z}} \text{Trace}(\Gamma^{(p)}\hat{\mathbf{Z}})\right)^{1/p} \quad \text{subject to } \mathbf{Z} \in \mathcal{S} \tag{7.8}$$

$$D_\infty(X,Y) = \frac{1}{2}\min_{\mathbf{Z}} \max_{i,i',j,j' \colon \mathbf{Z}\neq 0} \Gamma_{ij,i'j'} \quad \text{subject to } \mathbf{Z} \in \mathcal{S} \tag{7.9}$$

where $\mathcal{S} = \{\mathbf{Z} \in \mathbb{R}^{nm+1 \times nm+1} \colon \sum_i \mathbf{Z}_{ij,nm+1} = \lambda_j, \sum_j \mathbf{Z}_{ij,nm+1} = \nu_i, \mathbf{Z} = \mathbf{Z}^\top, \mathbf{Z} \geqslant 0, \text{rank}(\mathbf{Z}) = 1\}$ and $\Gamma^{(p)}$ denotes the p-th power of the matrix Γ entrywise.

The constraint $\text{rank}(\mathbf{Z}) = 1$ can be relaxed to the convex constraint $\mathbf{Z} \succeq 0$ (which means $\mathbf{Z}$ is symmetric and positive semidefinite) and additional linear constraints satisfied by the rank 1 matrix can be added to make the relaxation tighter.

Using this recipe we can construct the following family of semidefinite programming relaxations of the Gromov-Wasserstein and Gromov-Hausdorff distances.

$$\tilde{d}_{\mathcal{A},p}(X,Y) = \frac{1}{2}\left(\frac{1}{n^2}\min_{\mathbf{Z}} \text{Trace}(\Gamma^{(p)}\hat{\mathbf{Z}})\right)^{1/p} \quad \text{subject to } \mathbf{Z} \in \mathcal{A} \tag{7.10}$$

$$\tilde{d}_{\mathcal{A},\infty}(X,Y) = \frac{1}{2}\min_{\mathbf{Z}} \max_{i,j,i',j' \colon \mathbf{Z}\neq 0} \Gamma_{ij,i'j'} \quad \text{subject to } \mathbf{Z} \in \mathcal{A} \tag{7.11}$$

where we can consider different convex sets $\mathcal{A}$ as relaxing to different distances.

a. For a relaxation of the Gromov-Hausdorff distance (or Gromov-Wasserstein for uniform weights $\lambda_j = \nu_i = 1/\max\{n,m\}$ for all $j = 1,\ldots,m$ and

$i = 1, \ldots, n)^2$ consider

$$\mathcal{A} = \mathcal{GH}\colon = \{\mathbf{Z} \in \mathbb{R}^{nm+1 \times nm+1} : \sum_i \mathbf{Z}_{ij,nm+1} \geqslant 1, \sum_j \mathbf{Z}_{ij,nm+1} \geqslant 1,$$
$$\sum_{i,i'} \mathbf{Z}_{ij,i'j'} \geqslant 1, \sum_{j,j'} \mathbf{Z}_{ij,i'j'} \geqslant 1, \hat{\mathbf{Z}}\mathbf{1} = \max\{n, m\}\mathbf{z}, 0 \leqslant \mathbf{Z} \leqslant 1, \mathbf{Z} \succeq 0\}.$$

Relaxation (7.11) provides a lower bound for the Gromov-Hausdorff distance, since every element of $\mathcal{R}(X, Y)$ induces, up to normalization, a feasible $\mathbf{Z}$. In fact, if the optimal solution of equation (7.1) corresponds to $R \in \mathcal{R}(X, Y)$ such that $(x_i, y_j), (x_i, y_{j'}) \in R$ for some $j \neq j'$, the solution of equation (7.10) may split the mass in a way so $\mathbf{Z}_{ij,nm+1} + \mathbf{Z}_{ij',nm+1} = 1$ instead of having $\mathbf{Z}_{ij,nm+1} = \mathbf{Z}_{ij',nm+1} = 1$.

b. If $|X| = |Y|$ we may want to restrict the set of all correspondences between X and Y (where every element of X is related to at least one element in Y and vice versa) to the set of all bijective correspondences. In that case we can consider a tighter relaxation, that relaxes the registration problem and is similar to the one in [38].

$$\mathcal{A} = \text{Reg}\colon = \{\mathbf{Z} \in \mathbb{R}^{n^2+1 \times n^2+1}\colon \sum_{i=1}^{n} \mathbf{Z}_{ij,n^2+1} = 1, \sum_{j=1}^{n} \mathbf{Z}_{ij,n^2+1} = 1,$$
$$\mathbf{Z}_{n^2+1,n^2+1} = 1, \sum_{i,i'=1}^{n} \mathbf{Z}_{ij,i'j'} = 1, \sum_{j,j'=1}^{n} \mathbf{Z}_{ij,i'j'} = 1, \mathbf{Z}_{ij,ij'} = 0 \text{ if } j \neq j',$$
$$\mathbf{Z}_{ij,ij'} = 0 \text{ if } i \neq i', \text{Trace}(\mathbf{Z}) = n + 1, \hat{\mathbf{Z}}\mathbf{1} = n\mathbf{z}, 0 \leqslant \mathbf{Z} \leqslant 1, \mathbf{Z} \succeq 0\}$$

[2] By appropriately choosing the right hand side of the equality constraints in $\mathcal{GH}$ one can obtain a semidefinite relaxation of Gromov-Wasserstein distance for any weights.

c. The registration relaxation can be extended to finite metric spaces with different numbers of points. Let as before $|X| = n$ and $|Y| = m$. First, without loss of generality assume that $n \geqslant m$. Now consider the problem (7.1) where the set $\mathcal{R}(X, Y)$ is restricted to surjective functions $X \to Y$. Then relax the feasible set to the convex set:

$$\mathcal{A} = \text{Sur}\colon = \{\mathbf{Z} \in \mathbb{R}^{nm+1 \times nm+1}\colon \sum_{i=1}^{n} \mathbf{Z}_{ij,nm+1} \geqslant 1,\ \sum_{j=1}^{m} \mathbf{Z}_{ij,nm+1} = 1,$$
$$\mathbf{Z}_{nm+1,nm+1} = 1,\ \sum_{i,i'=1}^{n} \mathbf{Z}_{ij,i'j'} \geqslant 1,\ \sum_{j,j'=1}^{m} \mathbf{Z}_{ij,i'j'} = 1,$$
$$\mathbf{Z}_{ij,ij'} = 0 \text{ if } j \neq j',\ \text{Trace}(\mathbf{Z}) = n + 1,\ \hat{\mathbf{Z}}\mathbf{1} = n\mathbf{z},\ 0 \leqslant \mathbf{Z} \leqslant 1,\ \mathbf{Z} \succeq 0\}$$

Note that the set of constraints assumes that $i, i' = 1, \ldots n$, $j, j' = 1, \ldots m$ and $n \geqslant m$ and it is not symmetric with respect to i and j. Also note that under this relaxation, there exist sets that $d_{\mathcal{GH}}(X, Y) = 0$ but the relaxed distance (7.10) with $\mathcal{A} = \text{Sur}$ satisfies $\tilde{d}_{\text{Sur},p}(X, Y) \neq 0$ (and the same phenomena occurs for $\mathcal{A} = \text{Reg}$). For instance $X = \{x, x, y\}$ and $Y = \{x, y, y\}$. This is an artifact of only allowing surjective functions instead of all possible relations in $\mathcal{R}(X, Y)$.

Remark 7.1.1. Even though the max objective in equation (7.11) is convex, it is not smooth, which we observe to significantly hurt the performance of the numerical implementations. This is one reason to consider the p-norm approach to this relaxation and define the family of SDP relaxations (7.10) for $1 \leqslant p < \infty$.

Remark 7.1.2. Note that linear constraints in the sets $\mathcal{A}$ are not linearly independent and the extra variable $\mathbf{z}$ is redundant. However, one can easily choose alternative sets of constraints (even with fewer constraints or where the extra constraint is not redundant) with the same objectives in mind: (i) the relaxation is tight when the spaces are isometric (ii) the corresponding objective value satisfies the triangle inequality when $|X| = |Y|$.

We are primarily interested in the semidefinite programming relaxations of the Gromov-Hausdorff distance for finite metric spaces, namely $\tilde{d}_{\mathcal{A},p}$, $1 \leqslant p \leqslant \infty$ for $\mathcal{A} = \mathcal{GH}, \mathrm{Reg}, \mathrm{Sur}$. In figure 7.1, we extend a diagram of Mémoli's to situate the SDP relaxations we study in this thesis.

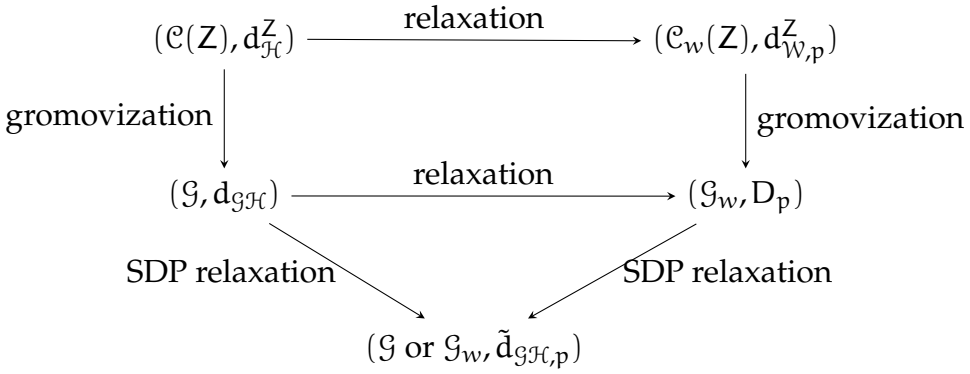


Figure 7.1: Diagram relating the different structures and distances
Here $\mathcal{G}$ is the collection of all compact metric spaces and $\mathcal{G}_w$ the collection of all metric measure spaces. The horizontal arrows represent the relaxation on the notion of correspondences. The gromovization arrows represent the process of getting rid of the ambient space.

7.2 Topological properties of the relaxed distances

In this section, we prove the main theoretical results. We begin by showing that the distances obtained by semidefinite relaxation are in fact pseudometrics on suitable subsets of the set of isometry classes of finite metric spaces; i.e., these distances satisfy all the axioms for a metric except that there exist distinct finite metric spaces such that the relaxed distance between them is 0. We then study various properties of the induced topology, proving analogues of the standard results about the topology induced on the set of isometry classes of compact metric spaces by the Gromov-Hausdorff distance.

7.2.1 Pseudometrics

Let $\mathcal{G}_{<\infty}$ the set of all finite metric spaces. First, we observe that $\tilde{d}_{\mathcal{A},\infty}$ is a pseudometric in $\mathcal{G}_{<\infty}$. However, if the spaces X, Y, Z have different numbers of points we cannot expect the triangle inequality to hold for $\tilde{d}_{\mathcal{A},p}$. That is because the triangle inequality does not even hold for tight solutions of equation (7.10) (i.e., rank 1 solutions, corresponding to elements of $\mathcal{R}(X,Y)$). This is an artifact of replacing the max with a sum.

In order to illustrate that fact we consider a simple example. Let $d_{\mathcal{GH},1}$ be the optimal of (7.10) for $p = 1$ and $\mathcal{A}$ the domain of (7.4) (i.e. the solutions corresponding to elements of $\mathcal{R}(X,Y)$). Then consider $X = \{x, y\}$, $Y = \{x, x, y\}$, $Z = \{y\}$, and observe that triangle inequality is not satisfied since $d_{\mathcal{GH},1}(X,Y) = 0$, $d_{\mathcal{GH},1}(Y,Z) = 2d(x,y)$ and $d_{\mathcal{GH},1}(X,Z) = d(x,y)$.

Nonetheless, if we consider the set of metric spaces with n points, which we denote by $\mathcal{G}_n$, we will show that $\tilde{d}_{\mathcal{A},p}$ for $1 \leqslant p < \infty$ is a pseudometric on $\mathcal{G}_n$. The most interesting part of this verification is the triangle inequality, which we prove in Theorem 7.2.1 below. In contrast to the situation with the Gromov-Hausdorff distance, passing to isometry classes of finite metric spaces does not suffice to produce an actual metric. Of course, if X and Y are isometric spaces then $\tilde{d}_{\mathcal{A},p}(X,Y) = 0$. By construction $\tilde{d}_{\mathcal{A},p}(\cdot,\cdot) \geqslant 0$ and the isometry between X and Y induces a feasible solution for equation (7.10) with objective value 0. However, there exists non-isometric spaces X, Y such that $\tilde{d}_{\mathcal{A},p}(X,Y) = 0$. Examples of that phenomenon can be constructed by observing that the graph isomorphism problem can be reduced to deciding whether the Gromov-Hausdorff distance is zero. Given a graph $G = (V, E)$ one then constructs a metric space $X(G)$ where

$$d(v,v') = \begin{cases} 1 & \text{if } (v,v') \in E \\ K \gg |V| & \text{otherwise.} \end{cases} \tag{7.12}$$

Therefore, given two graphs G, G' we have that G, G' are isomorphic if and only if $d_{\mathcal{GH}}(X(G), X(G')) = 0$. There exist explicit examples in the literature of graphs where any SDP relaxation on $|V|^2 \times |V|^2$ matrices cannot distinguish between two non-isomorphic graphs [53]. For such examples, $\tilde{d}_{\mathcal{A},p}(X(G), X(G')) = 0$ (see Figure 7.2).

Theorem 7.2.1. *Consider* $\tilde{d}_{\mathcal{A},p}$ *and* $\tilde{d}_{\mathcal{A},\infty}$ *defined in equations* (7.10) *and* (7.11) *respectively for* $\mathcal{A} = \mathcal{GH}, \mathrm{Reg}, \mathrm{Sur}$. *Then we have:*

a. For $X, Y, W \in \mathcal{G}_n$, *and* $1 \leqslant p < \infty$, $\tilde{d}_{\mathcal{A},p}$ *satisfies the triangle inequality.*

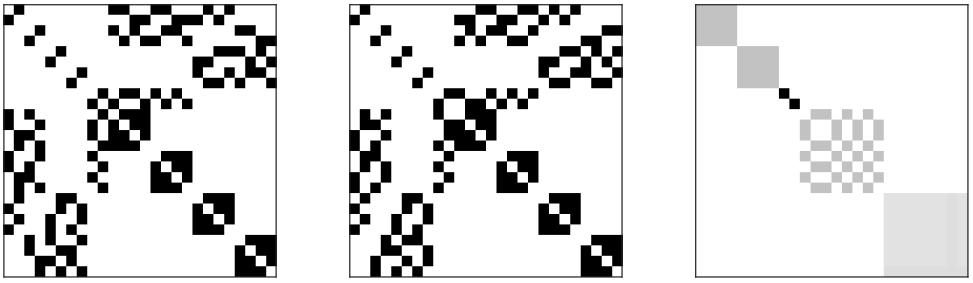


Figure 7.2: Two non-isometric metric spaces that have relaxed distance 0. We consider 3XOR instances with 5 variables and 4 equations and we construct the reduction from 3XOR to graph isomorphism from [53]. The left and middle figures represent corresponding adjacency graphs obtained after the reduction from the following system of equations in $\mathbb{Z}_2$:

$$\begin{aligned} x_1 \oplus x_2 \oplus x_5 &= b_1 \\ x_1 \oplus x_2 \oplus x_5 &= b_2 \\ x_1 \oplus x_3 \oplus x_4 &= b_3 \\ x_2 \oplus x_3 \oplus x_4 &= b_4 \end{aligned}, \text{ with } \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \text{ (left), and } \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \text{ (middle).}$$

Each graph has 26 vertices. We construct finite metric spaces X and Y according to (7.12) and we use SDPNAL+ [71] to compute the the relaxed distance, obtaining $\tilde{d}_{\mathrm{Reg},1}(X,Y) = 0$. The minimizer $\mathbf{Z}$ of (7.10) is rank 16. The figure in the right shows a *soft assignment* between X and Y obtained from $\mathbf{Z}$ by computing $\hat{\mathbf{Z}}\mathbf{1}$ and rearranging accordingly.

b. For $X, Y, W \in \mathcal{G}_{<\infty}$, $\tilde{d}_{\mathcal{A},\infty}$ satisfies the triangle inequality.

Proof. We begin by proving part (a). Note that it suffices to show that for $p \geqslant 1$,

$$\tilde{d}_{\mathcal{A},p}(X,W)^p \leqslant \tilde{d}_{\mathcal{A},p}(X,Y)^p + \tilde{d}_{\mathcal{A},p}(Y,W)^p. \tag{7.13}$$

This follows from the fact that for $a, b > 0$ and $p \geqslant 1$ we have $a^p + b^p \leqslant (a+b)^p$ and therefore if equation (7.13) holds we have:

$$\tilde{d}_{\mathcal{A},p}(X,W) \leqslant \sqrt[p]{\tilde{d}_{\mathcal{A},p}(X,Y)^p + \tilde{d}_{\mathcal{A},p}(Y,W)^p} \leqslant \tilde{d}_{\mathcal{A},p}(X,Y) + \tilde{d}_{\mathcal{A},p}(Y,W).$$

Now let $\mathbf{Z}$ and $\mathbf{V}$ the minimizers in equation (7.10) for X, Y and Y, W respectively in $\mathcal{A}$. From $\mathbf{Z}$ and $\mathbf{V}$ we construct $\mathbf{T}$ feasible for X, W in equation (7.10) and we show the objective function in $\mathbf{T}$ is smaller or equal to $\tilde{d}_{\mathcal{A},p}(X,Y) + \tilde{d}_{\mathcal{A},p}(Y,W)$.

If $x_i, x_{i'} \in X$, $y_j, y_{j'} \in Y$ and $w_k, w_{k'} \in W$ let $\mathbf{T}$ the unique feasible matrix in $\mathcal{A}$ that satisfies

$$\hat{\mathbf{T}}_{ik,i'k'} := \sum_{j,j'} \hat{\mathbf{Z}}_{ij,i'j'} \hat{\mathbf{V}}_{jk,j'k'}. \tag{7.14}$$

To see that $\mathbf{T}$ is well-defined, observe that it is straightforward to check that $\mathbf{T}$ satisfies the linear and inequality constraints of $\mathcal{A}$ using the fact that $\mathbf{Z}$ and $\mathbf{V}$ belong to $\mathcal{A}$. In order to verify that $\mathbf{T}$ is positive semidefinite, consider the Cholesky decompositions of $\mathbf{Z}$ and $\mathbf{V}$. Then

$$\hat{\mathbf{Z}}_{ij,i'j'} = z_{ij}^\top z_{i'j'}, \quad \hat{\mathbf{V}}_{jk,j'k'} = v_{jk}^\top v_{j'k'}$$

where z and ν do not necessarily correspond to the last column in equation (7.7). In fact z is a $r \times n^2$ matrix where r is the rank of $\hat{\mathbf{Z}}$ and z_{ij} is the column of z indexed by $i = 1, \ldots, n$ and $j = 1, \ldots, n$. Then note

$$\hat{\mathbf{T}}_{ik,i'k'} = \sum_{j,j'} \hat{\mathbf{Z}}_{ij,i'j'} \hat{\mathbf{V}}_{jk,j'k'} = \left\langle \sum_{j} z_{ij} \otimes \nu_{jk}, \sum_{j'} z_{i'j'} \otimes \nu_{j'k'} \right\rangle$$

therefore $\hat{\mathbf{T}}$ is PSD since it is a Gram matrix, and $\mathbf{T}$ is PSD since it has the same rank as $\hat{\mathbf{T}}$.

For the triangle inequality we need to show

$$\begin{aligned}\sum_{i,i'} \sum_{k,k'} \mathbf{T}_{ik,i'k'} |d_X(x_i, x_{i'}) - d_W(w_k, w_{k'})| \leqslant \\ \sum_{i,i'} \sum_{j,j'} \mathbf{Z}_{ij,i'j'} |d_X(x_i, x_{i'}) - d_Y(y_j, y_{j'})| \\ + \sum_{j,j'} \sum_{k,k'} \mathbf{V}_{jk,j'k'} |d_X(x_j, x_{j'}) - d_W(w_k, w_{k'})|. \quad (7.15)\end{aligned}$$

In the case we are dealing with, where $|X| = |Y| = |W|$, the constraints

$$\sum_{i,i'} \mathbf{Z}_{ij,i'j'} \geqslant 1 \qquad \text{and} \qquad \sum_{k,k'} \mathbf{V}_{jk,j'k'} \geqslant 1$$

are tight, meaning that equality holds, so for all j, j', we can multiply by 1 and rewrite the RHS of equation (7.15) as

$$\begin{aligned}&\sum_{i,i'} \sum_{j,j'} \mathbf{Z}_{ij,i'j'} |d_X(x_i, x_{i'}) - d_Y(y_j, y_{j'})| \sum_{k,k'} \mathbf{V}_{jk,j'k'} \\ &\qquad + \sum_{j,j'} \sum_{k,k'} \mathbf{V}_{jk,j'k'} |d_X(x_j, x_{j'}) - d_W(w_k, w_{k'})| \sum_{i,i'} \mathbf{Z}_{ij,i'j'} \\ &= \sum_{i,i'} \sum_{k,k'} \sum_{j,j'} \mathbf{Z}_{ij,i'j'} \mathbf{V}_{jk,j'k'} (|d_X(x_i, x_{i'}) - d_Y(y_j, y_{j'})| + |d_Y(y_j, y_{j'}) - d_W(w_k, w_{k'})|)\end{aligned}$$

Now it is clear that equation (7.15) follows from the triangle inequality in $\mathbb{R}$, which completes the verification of part (a).

In order to prove part (b), now we let X, Y, Z to be finite metric spaces with arbitrary number of points. And we let $\mathbf{Z}$ and $\mathbf{V}$ the minimizers in equations (7.11) for X, Y and Y, W respectively as before. We define $\mathbf{T}$ as in equation (7.14). We know $\mathbf{T}$ is feasible for equation (7.11) so the remaining step to prove is

$$\max_{\mathbf{T}\neq 0} \Gamma_{ik,i'k'} \leqslant \max_{\mathbf{Z}\neq 0} \Gamma_{ij,i'j'} + \max_{\mathbf{V}\neq 0} \Gamma_{jk,j'k'} \tag{7.16}$$

Let $(ik, i'k')$ the arg max of the left hand side of equation (7.16). Since $\mathbf{T}_{ik,i'k'} \neq 0$ and $\mathbf{T}_{ik,i'k'} = \sum_{j,j'} \mathbf{Z}_{ij,i'j'} \mathbf{T}_{kj,k'j'}$ then there exists j, j' such that $\mathbf{Z}_{ij,i'j'} \neq 0$ and $\mathbf{T}_{kj,k'j'} \neq 0$. Then we have

$$\begin{aligned}
\tilde{d}_{\mathcal{A},\infty}(X, W) &\leqslant \max_{\mathbf{T}\neq 0} \Gamma_{ik,i'k'} = |d_X(x_i, x_{i'}) - d_W(w_k, w_{k'})| \\
&\leqslant |d_X(x_i, x_{i'}) - d_Y(y_j, y_{j'})| + |d_Y(y_j, y_{j'}) - d_W(w_k, w_{k'})| \\
&\leqslant \max_{\mathbf{Z}\neq 0} \Gamma_{ij,i'j'} + \max_{\mathbf{V}\neq 0} \Gamma_{jk,j'k'} = \tilde{d}_{\mathcal{A},\infty}(X, Y) + \tilde{d}_{\mathcal{A},\infty}(Y, W). \quad (7.17)
\end{aligned}$$

□

Remark 7.2.1. The same argument will show that $\tilde{d}_{\mathcal{GW},p}$ satisfies triangle inequality as long as we add the constraint $\hat{\mathbf{Z}}\mathbf{1} = n\mathbf{z}$, where $n = |X| = |Y| = |W|$ and the measure of each of the points is equal $1/n$.

7.2.2 Monotonicity and continuity properties

The following lemma shows the monotonicity of $\tilde{d}_{\mathcal{A},p}$ with respect to p. The second part of the lemma proves continuity of $\tilde{d}_{\mathcal{A},p}$ at infinity.

Proposition 7.2.2. *For any X,Y finite metric spaces we have:*

a. If $1 \leqslant p \leqslant q < \infty$ *then* $\tilde{d}_{\mathcal{A},p}(X,Y) \leqslant \tilde{d}_{\mathcal{A},q}(X,Y) \leqslant \tilde{d}_{\mathcal{A},\infty}(X,Y)$.

b. $\lim_{p\to\infty} \tilde{d}_{\mathcal{A},p}(X,Y) = \min\limits_{\mathbf{Z}\in\tilde{\mathcal{T}}} \max\limits_{i,j,i',j':\, \mathbf{Z}\neq 0} \Gamma_{ij,i'j'} = \tilde{d}_{\mathcal{A},\infty}(X,Y)$.

Proof. Let $\mathbf{Z} \in \mathcal{A}$ optimal for equations (7.11) or (7.10) for some value of p. Then $\mathbf{1}^\top \hat{\mathbf{Z}} \mathbf{1} = n^2$. The weighted power mean inequality implies

$$\left(\frac{1}{n^2}\sum_{ij,i'j'} \Gamma^p_{ij,i'j'}\mathbf{Z}_{ij,i'j'}\right)^{1/p} \leqslant \left(\frac{1}{n^2}\sum_{ij,i'j'} \Gamma^q_{ij,i'j'}\mathbf{Z}_{ij,i'j'}\right)^{1/q} \leqslant \max_{i,j,i',j':\, \mathbf{Z}_{ij,i'j'}\neq 0} \Gamma_{ij,i'j'}$$

and taking the infimum in $\mathbf{Z}$ we obtain (a).

Now for fixed $\mathbf{Z}$ let $\Gamma^*_{\mathbf{Z}} = \max\{\Gamma_{ij,i'j'} \colon \mathbf{Z}_{ij,i'j'} \neq 0\}$, then using the standard calculus argument

$$\lim_{p\to\infty}\left(\frac{1}{n^2}\sum_{ij,i'j'} \Gamma^p_{ij,i'j'}\mathbf{Z}_{ij,i'j'}\right)^{1/p} = \Gamma^*_{\mathbf{Z}} \lim_{p\to\infty}\left(\sum_{ij,i'j'}\left(\frac{\Gamma_{ij,i'j'}}{\Gamma^*_{\mathbf{Z}}}\right)^p \frac{\mathbf{Z}_{ij,i'j'}}{n^2}\right)^{1/p} = \Gamma^*_{\mathbf{Z}}.$$

and taking infimum in $\mathbf{Z}$ we obtain (b). □

Proposition 7.2.2 holds for metric spaces X and Y with possibly different number of points and it says that even though $\tilde{d}_{\mathcal{A},p}$ may not satisfy the triangle inequality, it does in the limit $p \to \infty$.

7.2.3 Extension of the distance to compact infinite sets

Every compact metric space X is the limit of a sequence of finite metric spaces in the Gromov-Hausdorff topology, denoted here by $\tau_{\mathcal{GH}}$ (see for instance [19, Example 7.4.9]). In fact, by taking $\epsilon_n \to 0$ and choosing a finite ϵ_n-net S_n in X for every n, we get $S_n \overset{\mathcal{GH}}{\to} X$ because

$$d_{\mathcal{GH}}(X, Sn) \leqslant d_{\mathcal{H}}(X, S_n) \leqslant \epsilon_n.$$

This property inspires the following definition of an actual distance between compact metric spaces.

Definition 7.2.1. Let X, Y compact metric spaces. Given $\epsilon_n \to 0$, let X_n, Y_n respective ϵ_n-nets of X and Y, with the same number of points N. Define

$$\hat{d}_{\mathcal{A},p}(X, Y) = \inf_{\epsilon_n,\, X_n\, Y_n} \limsup_{n\to\infty} \tilde{d}_{\mathcal{A},p}(X_n, Y_n) \tag{7.18}$$

Note that lim sup exists because for all n we have

$$\tilde{d}_{\mathcal{A},p}(X_n, Y_n) \leqslant \tilde{d}_{\mathcal{A},\infty}(X_n, Y_n) \leqslant \frac{1}{2}\max(\mathrm{diam}(X), \mathrm{diam}(Y)).$$

Also, note the triangle inequality holds for this limit, which also implies that $\hat{d}_{\mathcal{A},p}$ and $\tilde{d}_{\mathcal{A},p}$ may not agree.

To illustrate how the right hand side of (7.18) behaves, let's say that $|X| < |Y|$ and for some n the ϵ_n-net Y_n of Y has at least N points and $|X| < N$, then consider X_n to be X with some repeated points and run the SDP (7.10) or (7.11) to compute $\tilde{d}_{\mathcal{A},p}(X_n, Y_n)$ so that $|Y_n| = |X_n| = N$. Note that this is well

defined and when $\tilde{d}_{\mathcal{A},p}(X_n, X)$ exists (i.e. $\mathcal{A} = \mathcal{GH}, \mathrm{Sur}$) we have $\tilde{d}_{\mathcal{A},p}(X_n, X) = 0$ because the matrix $\mathbf{Z}$ corresponding to the surjective function $X_n \to X$ is in $\mathcal{A}$ and has objective value 0.

7.2.4 Comparison with the Gromov-Hausdorff distance

Let $\mathcal{X}$ denote the set of isometry classes of compact metric spaces. Definition 7.2.1 extends the relaxed distances (7.10) and (7.11) to $\mathcal{X}$, obtaining the function $\hat{d}_{\mathcal{A},p} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Lemma 7.2.3. *For $X, Y \in \mathcal{X}$ we have for $1 \leqslant p \leqslant \infty$*

$$\hat{d}_{\mathcal{A},p}(X, Y) \leqslant d_{\mathcal{GH}}(X, Y)$$

Proof. First assume X, Y are finite and let $R \in \mathcal{R}(X, Y)$ the minimizer in (7.1). If $|R| = N$ let X_N, Y_N the ϵ-net so that every element of X appears in X_N as many times as it appears in R (and the same for Y_N). Then the bijective function between X_N and Y_N corresponding to R induces a feasible $\mathbf{Z}$, proving the result in the finite case. For the remaining case consider a ϵ_n-net where $\epsilon_n \to 0$. □

Now consider X, Y finite metric spaces. First observe that $\mathcal{A} = \mathcal{GH}$ includes the $\mathbf{Z}$ induced by all elements in $\mathcal{R}(X, Y)$, which together with Proposition 7.2.2 implies

$$\tilde{d}_{\mathcal{GH},p}(X, Y) \leqslant \tilde{d}_{\mathcal{GH},\infty}(X, Y) \leqslant d_{\mathcal{GH}}(X, Y).$$

Since $\mathrm{Sur} \subset \mathcal{GH}$ we have

$$\tilde{d}_{\mathcal{GH},p}(X,Y) \leqslant \tilde{d}_{\mathrm{Sur},p}(X,Y),$$

and if $|X| = |Y|$ we can consider that $\mathrm{Reg} \subset \mathrm{Sur}$ therefore

$$\tilde{d}_{\mathcal{GH},p}(X,Y) \leqslant \tilde{d}_{\mathrm{Sur},p}(X,Y) \leqslant \tilde{d}_{\mathrm{Reg},p}(X,Y).$$

Also, the smaller the set $\mathcal{A}$, the more likely is the relaxation to produce a tight solution (a rank 1 solution, corresponding with an element of $\mathcal{R}(X,Y)$). Note that neither $\tilde{d}_{\mathrm{Sur},p}$ nor $\tilde{d}_{\mathrm{Reg},p}$ are comparable with $d_{\mathcal{GH}}$.

7.2.5 Topologies induced by relaxed distances

Any metric or pseudometric d defines a topology τ characterized by the property that given a sequence X_n, we have convergence $X_n \xrightarrow{\tau} X$ if and only if $d(X_n, X) \to 0$. In particular, the Gromov-Hausdorff distance induces a topology on the set of isometry classes of compact metric spaces. The Gromov-Hausdorff topology is a fairly weak topology; for example, there are many compact sets. Proposition 7.4.12 in [19] characterizes the Gromov-Hausdorff convergence in terms of ϵ-nets, implying that if a sequence $\{X_n\}$ converges in the Gromov-Hausdorff topology, then for all $\epsilon > 0$, the cardinality of ϵ-nets is uniformly bounded over X_n. Therefore if a class $\mathcal{X}$ of metric spaces is pre-compact (i.e. any sequence of elements of $\mathcal{X}$ has a convergent subsequence) in the Gromov-Hausdorff topology, then for every $\epsilon > 0$ the size of a minimal ϵ-net is uniformly bounded over all elements of $\mathcal{X}$. The analysis in [19] shows that this property of $\mathcal{X}$, along with the fact

that the diameters of its members are uniformly bounded (what is called totally boundedness, Definition 7.2.2), is sufficient for pre-compactness (Theorem 7.2.5).

Let $\hat{\tau}_{\mathcal{A},p}$ $(1 \leqslant p \leqslant \infty)$, and $\tilde{\tau}_{\mathcal{A},\infty}$ denote the topologies induced by the pseudometrics $\hat{d}_{\mathcal{A},p}$ $(1 \leqslant p \leqslant \infty)$ and $\tilde{d}_{\mathcal{A},\infty}$, respectively. We obtain an analogous characterization of pre-compact sets in the topology for $\hat{\tau}_{\mathcal{A},p}$ for $1 \leqslant p \leqslant \infty$ in Corollary 7.2.6 below.

Proposition 7.2.4. *(Proposition 7.4.12 in [19]) For compact metric spaces* X *and* $\{X_n\}_{n=1}^{\infty}$, $X_n \xrightarrow{\tau_{\mathcal{GH}}} X$ *if and only if the following holds. For every* $\epsilon > 0$ *there exist a finite* ϵ*-net* S *in* X *and an* ϵ*-net* S_n *in each* X_n *such that* $S_n \xrightarrow{\tau_{\mathcal{GH}}} S$*. Moreover these* ϵ*-nets can be chosen so that, for all sufficiently large* n*,* S_n *have the same cardinality as* S*.*

Note that by construction (Definition 7.2.1) the characterization of convergence by ϵ-nets from Proposition 7.2.4 is also true when substituting $\tau_{\mathcal{GH}}$ by $\hat{\tau}_{\mathcal{A},p}$, $1 \leqslant p \leqslant \infty$.

Definition 7.2.2. (Definition 7.4.14 in [19]) A class $\mathcal{X}$ of compact metric spaces is totally bounded if

a. There exists a constant D such that for all $X \in \mathcal{X}$, $\operatorname{diam}(X) < D$.

b. For every $\epsilon > 0$ there exists a number N_ϵ such that every $X \in \mathcal{X}$ contains an ϵ-net consisting of at most N_ϵ points.

Theorem 7.2.5. *(Theorem 7.4.15 in [19]) Any uniformly totally bounded class $\mathcal{X}$ of compact metric spaces is pre-compact in the Gromov-Hausdorff topology.*

By Theorem 7.2.5, we know that if $\mathcal{X}$ is totally bounded and $\{X_n\}_{n=1}^{\infty}$ is a sequence in $\mathcal{X}$ then it contains a convergent subsequence in $\mathcal{X}$. Since $\hat{d}_{\mathcal{A},p} \leqslant d_{\mathcal{GH}}$, that subsequence is also convergent in $\hat{\tau}_{\mathcal{A},p}$, which immediately implies the following corollary:

Corollary 7.2.6. *Any uniformly totally bounded class $\mathcal{X}$ of compact metric spaces is pre-compact in the topology $\hat{\tau}_{\mathcal{A},p}$ for $1 \leqslant p \leqslant \infty$.*

7.2.6 Local topological properties

In the space of compact metric spaces we know $d_{\mathcal{GH}}(X,Y) = 0$ if and only if X and Y are isometric. The example at the beginning of Section 7.2.1 shows that this is not true for $\tilde{d}_{\mathcal{A},p}$ in general. However, in this section we show it is true for *most* finite X.

Definition 7.2.3. Let X a finite metric space. We say that X is generic if $X \in \mathcal{G}_{<\infty}$ and all pairwise distances in X are different and non-zero.

The name generic is justified in the following sense: if $X \in \mathcal{G}_n$ is not generic, for all $\epsilon > 0$ there exists $Y \in \mathcal{G}_n$ such that $d_{\mathcal{GH}}(X,Y) < \epsilon$ and Y is generic. Also, if $X \in \mathcal{G}_n$ is generic there exists $\epsilon > 0$ such that for all $Y \in \mathcal{G}_n$ that satisfy $d_{\mathcal{GH}}(X,Y) < \epsilon$ we have that Y is also generic. Which proves the following remark:

Remark 7.2.2. The set of generic metric spaces is dense in $\tau_{\mathcal{GH}}|_{\mathcal{G}_{<\infty}}$ and open in $\tau_{\mathcal{GH}}|_{\mathcal{G}_n}$.

Lemma 7.2.7. *If* X *and* Y *are generic and* $\tilde{d}_{\mathcal{A},p}(X,Y) = 0$ *then* X *and* Y *are isometric.*

Proof. Assume without loss of generality $|X| \geqslant |Y|$ and $\tilde{d}_{\mathcal{A},p}(X,Y) = 0$. Let $\mathbf{Z}$ the solution of (7.10) for X, Y with objective value 0. Note that the constraint $\sum_{j,j'} \mathbf{Z}_{ij,i'j'} = 1$ for all i, i' implies that, given $i \neq i'$ there exists j, j' such that $\mathbf{Z}_{ij,i'j'} > 0$. Since the objective value is 0, that implies that $d_X(x_i, x_{i'}) = d_Y(y_j, y_{j'})$. Since all pairwise distances in X are different, that completely determines all distances in Y and in particular it implies $|X| = |Y|$, X and Y are isometric, and the unique solution of (7.10) corresponds to the isometry between X and Y. □

Corollary 7.2.8. *If* $X \in \mathcal{G}_n$ *is generic there exists a neighborhood of* X *in* $\mathcal{G}_n$ *such that for all* Y *in that neighborhood*

$$d_{\mathcal{GH}}(X,Y) = \frac{1}{2} \max_{i,j=1\ldots n} |d_X(x_i, x_j) - d_Y(y_i, y_j)|$$

(Y *is a small enough perturbation of a metric space isometric with* X *where we label the points such that the isometry is* $x_i \mapsto y_i$ *for all* i*). In particular we can think of the neighborhood where that property holds as the neighborhood of* X *with radius* Δ/n *where* Δ *is the smallest non-zero entry of the matrix* $\Gamma(X,X)$*. In this setting we have that*

$$\tilde{d}_{\mathcal{A},p}(X,Y) = \frac{1}{2} \left(\frac{1}{n^2} \sum_{i,j} |d_X(x_i, x_j) - d_Y(y_i, y_j)|^p \right)^{1/p}$$

in the neighborhood of X *of radius* Δ^p/n. *And, in the neighborhood of* X *of radius* Δ *we have*

$$\tilde{d}_{\mathcal{A},\infty}(X,Y) = d_{\mathcal{GH}}(X,Y).$$

This implies that the topologies $\tau_{\mathcal{GH}}|_{\mathcal{G}_n}$ *and* $\tilde{\tau}_{\mathcal{GH},p}|_{\mathcal{G}_n}$ *are equivalent for all finite* n *and* p. *And we have* $\tilde{d}_{\mathcal{A},\infty}$ *and* $d_{\mathcal{GH}}$ *are generically locally the same.*

Corollary 7.2.8 says that the metrics $\tilde{d}_{\mathcal{A},\infty}$ and $d_{\mathcal{GH}}$ are locally the same (while $\tilde{d}_{\mathcal{A},p}$ and $d_{\mathcal{GH}}$ are locally equivalent), whereas the discussion in Section 7.2.1 implies that $\tilde{d}_{\mathcal{A},\infty}$ and $d_{\mathcal{GH}}$ are not globally the same. Since the metric $d_{\mathcal{GH}}$ is intrinsic in $\mathcal{X}$ [35] (i.e.: the distance between two points coincide with the infimum of the lengths of path between the two points), then it is implied that neither $\tilde{d}_{\mathcal{A},\infty}$ nor $\tilde{d}_{\mathcal{A},p}$ are intrinsic metrics.

7.3 GHMatch: a rank-1 augmented Lagrangian approach towards the registration problem

In the previous sections, we have studied an approach to the problem of computing the Gromov-Hausdorff distance (equation (7.4)) via semidefinite optimization. Here we first lift the variable $\mu \in \mathbb{R}^{nm}$ to a symmetric variable $\mathbf{Z} \in \mathbb{R}^{nm+1 \times nm+1}$ such that $\text{rank}(\mathbf{Z}) = 1$. We then relax the non-convex rank constraint to the convex constraint $\mathbf{Z} \succeq 0$.

There are many attractive properties of the semidefinite relaxations. For one thing, there are many software packages that efficiently provide global solutions to semidefinite programs (e.g., SDPNAL+ [71]). Moreover,

there is a great deal of research energy directed at producing more efficient SDP solvers; the field is rapidly evolving and solvers are getting more efficient every day. Furthermore, SDP relaxations have the advantage of often being tight: in our situation, we have observed numerically that the solution $\mathbf{Z}$ frequently has rank 1. In this case, the semidefinite optimization finds the global solution of the original problem, and also provides a certificate of its optimality. This property has recently been exploited to efficiently produce certificates of optimality of solutions found by fast non-convex algorithms that typically may converge to local optima [9, 34].

On the other hand, the semidefinite relaxations have the disadvantage that they square the number of variables of the original problem: even with efficient solvers, this expansion makes these problems intractable for large sets of points. Also, when the SDP is not tight, it may produce a high rank solution $\mathbf{Z}$ that may not be easily rounded to a feasible μ.

Motivated by these concerns, in this section we propose a non-convex optimization approach for the registration problem. Here we trade the global optimality guarantee and the pseudometric the semidefinite optimization provides for computational efficiency and a guarantee of a true (albeit not necessarily globally optimal) correspondence.

We will assume that $|X| = |Y| = n$. By restricting equation (7.4) to this case and replacing the infinity norm by the p-norm formulation we obtain the following non-convex optimization problem, where $\mathbf{y} \in \mathbb{R}^{n^2}$ is indexed

by a pair of variables ij where $i,j = 1,\ldots,n$.

$$\min_{\mathbf{y}} \langle \Gamma^{(p)}, \mathbf{y}\mathbf{y}^\top \rangle \quad \text{subject to } \sum_{i=1}^{n} \mathbf{y}_{ij} = 1,\ \sum_{j=1}^{n} \mathbf{y}_{ij} = 1,\ 0 \leqslant \mathbf{y}_{ij} \leqslant 1 \tag{7.19}$$

Now, instead of considering a semidefinite relaxation as we did previously, we propose a greedy method to directly solve (7.19). Let $A \in \mathbb{R}^{2n \times n^2}$ such that $A\mathbf{y} = \mathbf{1}$ if and only if $\sum_{i=1}^{n} \mathbf{y}_{ij} = 1$ and $\sum_{j=1}^{n} \mathbf{y}_{ij} = 1$ and let $b = \mathbf{1} \in \mathbb{R}^{2n}$. Then equation (7.19) is equivalent to the following quadratic optimization problem with linear and box constraints.

$$\min_{\mathbf{y}} \mathbf{y}^\top \Gamma^{(p)} \mathbf{y} \quad \text{subject to } A\mathbf{y} = b,\ 0 \leqslant \mathbf{y} \leqslant 1 \tag{7.20}$$

In order to solve problem (7.20), we use a projected augmented Lagrangian approach (e.g., see [52, Algorithm 17.4]).

$$\mathcal{L}(\mathbf{y}, \lambda, \sigma) = \mathbf{y}^\top \Gamma^{(p)} \mathbf{y} - \lambda^\top (A\mathbf{y} - b) + \frac{\sigma}{2} \|A\mathbf{y} - b\|_2^2 \tag{7.21}$$

We propose the algorithm GHMatch (see Algorithm 3). Theoretical convergence analysis for GHMatch is left for future work. In the next section, we describe numerical experiments that indicate the performance of the algorithm. In the experiments we conducted, we found that this procedure converges to a local minimum of equation (7.19). That solution may be rounded to an actual correspondence between the point sets $\bar{\mathbf{y}}$, and therefore the value $\langle \Gamma, \bar{\mathbf{y}}\bar{\mathbf{y}}^\top \rangle$ is an upper bound for the Gromov-Hausdorff distance.

Algorithm 3: GHMatch

$\mathbf{y}_0 \leftarrow \frac{1}{n}\mathbf{1} \in \mathbb{R}^{n^2}$;
$\lambda_0 \leftarrow \mathbf{1} \in \mathbb{R}^{2n}$;
$\sigma_0 \leftarrow 5$;
$\mu \leftarrow 10$;
for $k = 0, 1, 2, \ldots$ **do**
    $\mathbf{y} \leftarrow \arg\min_{0 \leqslant \mathbf{y} \leqslant 1} \mathcal{L}(\mathbf{y}, \lambda_k, \sigma_k)$; // Use $\mathbf{y}_k$ as initial point for this minimization
    $\lambda_{k+1} \leftarrow \lambda_k - \sigma_k(A\mathbf{y}_{k+1} - b)$;
    $\sigma_{k+1} \leftarrow \mu\sigma_k$;
for $i = 1, \ldots, n$; // To find the map corresponding with $\mathbf{y}_T$
do
    $\mathrm{map}(i) = \arg\max_{j=1,\ldots,n} \mathbf{y}_T(1 + (i-1)n : in)$;

7.4 Numerical performance

In this section, we describe the results of a number of numerical experiments to explore the applications of our new distance and the performance of our augmented Lagrangian approach.

7.4.1 Classification using the distance $\tilde{d}_{\mathcal{GH}}$

In order to validate our distance numerically we compare with the numerical experiments described in [17], using data and algorithms available on Yaron Lipman's personal website [41]. As we describe below, we find that our procedure produces results that are competitive with this procedure.

In [17] the authors propose an algorithm to automatically quantify

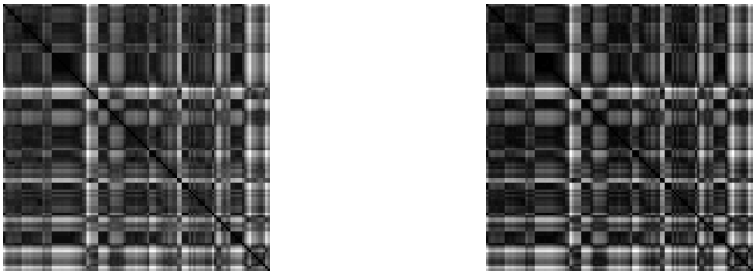


Figure 7.3: Visual comparison between $d_{\mathcal{GH}}$ and $\tilde{d}_{\mathcal{GH},1}$ on a real data set. (Left) Distance matrix obtained from computing the best rigid transformation that maps the corresponding labeled landmarks from the teeth dataset described in Section 7.4.1. (Right) Distance matrix obtained from computing the SDP distance $\tilde{d}_{\mathcal{GH},p}$ with $p = 1$. Darker color corresponds to smaller distance. We observe the same distance patterns in both matrices even though the scales are different.

the geometric similarity of anatomical surfaces based on a distance inspired by the Gromov-Wasserstein distance which is invariant under conformal maps. They experiment with a real dataset coming from surfaces of teeth of different species; they compare the results of their algorithm with a method based on an expert selecting 16 landmarks on each tooth and then finding the best rigid transformation to match the labeled landmarks. Specifically, they work with a dataset consisting of 116 teeth. For each tooth, they find the closest tooth according to each distance, and then see whether they are in the same category.

We perform the same experiment on 115 of the teeth (since one of them seems to be in a different scale), but without providing our algorithm with the correspondence between the landmarks. To be precise, we consider the metric spaces

$$X_i = \{p_1^i, \dots p_{16}^i\}, \quad i = 1 \dots 115.$$

The points of these metric spaces are the landmarks chosen by the expert, and the metric is given by the euclidean distance between the landmarks. We compare the distance matrix $\tilde{d}_{\mathcal{GH},p}(X_i, X_j)$ with the distance obtained by the software from [41] that finds the best rigid transformation that sends the n-th point of X_i to the n-th point of X_j for $n = 1, \ldots 16$. See Figure 7.3 for a visual comparison of the distance matrices.

When running the nearest-neighbor classification test as described above, we obtain very similar performance: 0.85 frequency of success in our distance against 0.91 for the conformal Wasserstein distance proposed in [17] and 0.92 for the landmark comparison algorithm that uses the a priori known correspondence. We find this result very encouraging given that our algorithm does not make any geometric assumptions about the teeth (e.g., we do not assume they are smooth surfaces), in contrast to [17].

7.4.2 Performance of GHMatch

In order to evaluate our non-convex optimization formulation of the registration problem, we consider the 3D models from [18] and we sample random points from each model. We run the rank 1 augmented lagrangian optimization from Algorithm 3, using Matlab's implementation of the reflective trust region algorithm to run the step

$$\mathbf{y} \leftarrow \arg\min_{0 \leqslant \mathbf{y} \leqslant 1} \mathcal{L}(\mathbf{y}, \lambda_k, \sigma_k).$$

In Figure 7.5 we depict the resulting map between corresponding figures.

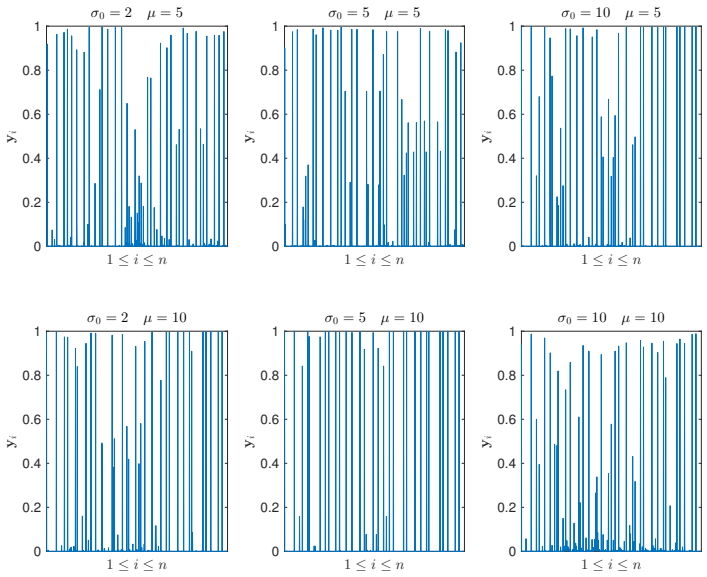


Figure 7.4: Convergence of GHMatch

This is the value of convergence of $\mathbf{y}_k$ for different parameters σ_0 and μ, for $n = 30$. All these $\mathbf{y}$ satisfy the linear constraint $A\mathbf{y} = b$ but the choice of the parameters determine how close the vector $\mathbf{y}$ is to a feasible vector with entries in $\{0, 1\}$.

Figure 7.5: Matching surfaces with GHMatch
We run GHMatch on 50 points sampled at random from the surfaces, and 60 points for the dogs. The pairwise distances we consider in each figure are the geodesic distances in the mesh. Images that are horizontally aligned correspond to different angles of the same correspondence between the 3D models. Note that for the dogs, the correspondence matches the left legs of one dog with the right legs of the other one (this is a consequence of the symmetry). Also note that there are small imperfections, like the tail of one dog matching with one leg of the other one (this is a consequence of randomly sampling and obtaining different number of points from different dogs tails). The algorithm with 50 sample points runs in less than 6 minutes on a standard 2013 MacBook Air.

By design we know $\sigma_k \to \infty$ as k increases, which guarantees that $\|A\mathbf{y}_k - b\| \to 0$. However, there is no theoretical guarantee that $\mathbf{y}_k$ will converge to a sparse vector with entries in $\{0, 1\}$. However, we have observed in our numerical simulations that $\mathbf{y}_k$ converges to a fairly sparse vector where most of the entries are close to 0 or 1 provided a good choice of the parameters μ and σ_0 (see Figure 7.4). Moreover, regardless of the choice of the parameters, we find that our thresholding step in the algorithm often obtains a map that is bijective.

Remark 7.4.1. As an alternative to the selection of parameters σ_0 and λ, a thresholding step could be introduced inside the main iteration (lines 5 to 9) to enforce sparsity of the resulting $\mathbf{y}$.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

[3] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 670–688, 2015.

[4] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[5] Yonathan Aflalo, Anastasia Dubrovina, and Ron Kimmel. Spectral generalized multi-dimensional scaling. *International Journal of Computer Vision*, 118(3):380–392, 2016.

[6] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine learning*, 75(2):245–248, 2009.

[7] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007.

[8] Pranjal Awasthi, Afonso S. Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, and Rachel Ward. Relax, no need to round: Integrality of clustering formulations. *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 191–200, 2015.

[9] Afonso S Bandeira. A note on probably certifiably correct algorithms. *Comptes Rendus Mathematique, to appear*, 2015.

[10] Afonso S Bandeira. Ten lectures and forty-two open problems in the mathematics of data science, 2015.

[11] Alexander I. Barvinok. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13(2):189–202, 1995.

[12] Shai Ben-David. Clustering is easy when ....what? *arXiv:1510.05336*, 2015.

[13] N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *arXiv:1605.08101*, 2016.

[14] Nicolas Boumal, P-A Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *arXiv:1605.08101*, 2016.

[15] Nicolas Boumal, Vladislav Voroninski, and Afonso S Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Neural Information Processing Systems (NIPS 2016)*, 2016.

[16] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[17] Doug M. Boyer, Yaron Lipman, Elizabeth St. Clair, Jesus Puente, Biren A. Patel, Thomas Funkhouser, Jukka Jernvall, and Ingrid Daubechies. Algorithms to automatically quantify the geometric similarity of anatomical surfaces. *Proceedings of the National Academy of Sciences*, 108(45):18221–18226, 2011.

[18] Alexander Bronstein, Michael Bronstein, and Ron Kimmel. *Numerical Geometry of Non-Rigid Shapes*. Springer Publishing Company, Incorporated, 1 edition, 2008.

[19] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Society Providence, 2001.

[20] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[21] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[22] Connor Clark and Jugal Kalita. A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics*, 2014.

[23] Hongbo Dong and Kurt Anstreicher. Separating doubly nonnegative and completely positive matrices. *Mathematical Programming*, pages 1–23, 2013.

[24] Nadav Dym and Yaron Lipman. Exact recovery with symmetries for procrustes matching. *arXiv:1606.01548*, 2016.

[25] Asi Elad (Elbaz) and Ron Kimmel. On bending invariant signatures for surfaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(10):1285–1295, 2003.

[26] Gábor Elek. Samplings and observables. Limits of metric measure spaces. arXiv:1205.6936, 2012.

[27] Ehsan Elhamifar, Guillermo Sapiro, and René Vidal. Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. *Advances in Neural Information Processing Systems*, pages 19–27, 2012.

[28] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[29] Michael Grant, Stephen Boyd, and Yinyu Ye. CVX: matlab software for disciplined convex programming, 2008.

[30] Mikhail Gromov. Groups of polynomial growth and expanding maps. *Inst. Hautes Études Sci. Publ. Math.*, (53):53–73, 1981.

[31] Mikhail Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Progress in Mathematics. Birkhäuser, 2001.

[32] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.

[33] Takayuki Iguchi, Dustin G. Mixon, Jesse Peterson, and Soledad Villar. On the tightness of an SDP relaxation of k-means. *arXiv:1505.04778*, 2015.

[34] Takayuki Iguchi, Dustin G. Mixon, Jesse Peterson, and Soledad Villar. Probably certifiably correct k-means clustering. *Mathematical Programming (to appear)*, pages 1–38, 2016.

[35] Alexandr Olegovich Ivanov, NK Nikolaeva, and Alexey Avgustinovich Tuzhilin. The gromov–hausdorff metric on the space of compact metric spaces is strictly intrinsic. *Mathematical Notes*, 100(5-6):883–885, 2016.

[36] Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 731–740. ACM, 2002.

[37] Michel Journée, Francis Bach, P-A Absil, and Rodolphe Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

[38] Itay Kezurer, Shahar Z. Kovalsky, Ronen Basri, and Yaron Lipman. Tight relaxation of quadratic matching. *Computer Graphics Forum*, 34(5):115–128, 2015.

[39] Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[40] Yann LeCun and Corinna Cortes. Mnist handwritten digit database. *AT&T Labs [Online]. http://yann.lecun.com/exdb/mnist visited April 2017*, 2010.

[41] Yaron Lipman. Software and teeth dataset. *http://www.wisdom.weizmann.ac.il/~ylipman/CPsurfcomp/ visited April 2017*, 2011.

[42] Anna Little and Alicia Byrd. A multiscale spectral method for learning number of clusters. In *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on*, pages 457–460. IEEE, 2015.

[43] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, 28(2):129–137, 1982.

[44] John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Ann. of Math. (2)*, 169(3):903–991, 2009.

[45] Haggai Maron, Nadav Dym, Itay Kezurer, Shahar Kovalsky, and Yaron Lipman. Point registration via efficient convex relaxation. *ACM Trans. Graph.*, 35(4):73:1–73:12, July 2016.

[46] Facundo Mémoli. On the use of Gromov-Hausdorff Distances for Shape Comparison. pages 81–90, Prague, Czech Republic, 2007. Eurographics Association.

[47] Facundo Mémoli. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, pages 1–71, 2011. 10.1007/s10208-011-9093-5.

[48] Facundo Mémoli and Guillermo Sapiro. A theoretical and computational framework for isometry invariant recognition of point cloud data. *Found. Comput. Math.*, 5(3):313–347, 2005.

[49] Dustin G. Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and inference (to appear)*, February 2016.

[50] Abhinav Nellore and Rachel Ward. Recovery guarantees for exemplar-based clustering. *Information and Computation*, 245:165–180, 2015.

[51] Yurii Nesterov, Arkadii Nemirovskii, and Yinyu Ye. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

[52] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

[53] Ryan O'Donnell, John Wright, Chenggang Wu, and Yuan Zhou. Hardness of robust graph isomorphism, lasserre gaps, and asymmetry of random graphs. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1659–1677. SIAM, 2014.

[54] Rafail Ostrovsky, Yuval Rabani, Leonard J Schulman, and Chaitanya Swamy. The effectiveness of Lloyd-type methods for the k-means

problem. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, 2006.

[55] Gábor Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research*, 23(2):339–358, 1998.

[56] Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18(1):186–205, 2007.

[57] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *Proc. ICML'16*, pages 2664–2672, 2016.

[58] Alexander Shapiro. Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika*, 47(2):187–199, 1982.

[59] Karl-Theodor Sturm. On the geometry of metric measure spaces. I. *Acta Math.*, 196(1):65–131, 2006.

[60] Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[61] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[62] Samuel Vaiter, Gabriel Peyré, and Jalal M Fadili. Model consistency of partly smooth regularizers. Technical report, Preprint Hal-00987293, 2014.

[63] Vijay V Vazirani. *Approximation algorithms*. Springer Science & Business Media, 2013.

[64] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications, Y. Eldar, G. Kutyniok (eds.), Cambridge University Press*, 2012.

[65] Cédric Villani. *Topics in optimal transportation*. Graduate studies in mathematics. American Mathematical Society, cop., Providence (R.I.), 2003.

[66] Soledad Villar. Clustering visualization. `http://solevillar.github.io/2016/07/05/Clustering-MNIST-SDP.html`, 2016. Blog entry in Todo epsilon suma.

[67] Soledad Villar. Implementation of the k-means semidefinite program, 2016. Software available from github.com.

[68] Soledad Villar, Afonso S. Bandeira, Andrew J. Blumberg, and Rachel Ward. A polynomial time relaxation for the gromov-hausdorff distance. *arXiv:1610.05214*, 2016.

[69] Ramya Korlakai Vinayak and Babak Hassibi. Similarity clustering in the presence of outliers: Exact recovery via convex program. In *IEEE*

International Symposium on Information Theory, ISIT 2016, Barcelona, Spain, July 10-15, 2016, pages 91–95, 2016.

[70] Bowei Yan and Purnamrita Sarkar. On robustness of kernel clustering. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3098–3106. Curran Associates, Inc., 2016.

[71] Liuqin Yang, Defeng Sun, and Kim-Chuan Toh. SDPnal+: a majorized semismooth newton-cg augmented lagrangian method for semidefinite programming with nonnegative constraints. *Mathematical Programming Computation*, 7(3):331–366, 2015.

Vita

María Soledad Villar Lozano was born in Montevideo, Uruguay, daughter of María del Pilar Lozano and José Enrique Villar. Her interest in Mathematics started at the age of 12 thanks to the Uruguay Mathematical Olympiads. She received the undergraduate degrees of *Licenciada en Matemática* from *Universidad de la República Oriental del Uruguay*, and *Ingeniera en Informática* from *Universidad Católica del Uruguay* in 2010. She was awarded a Fellowship from the Uruguayan National Research Agency (ANII) to pursue a master's degree in number theory and applications under the supervision of Dr. Gonzalo Tornaría. In 2012 she defended her master thesis *La fórmula de Gross sobre alturas y valores especiales de* L*-series*. Soledad entered the University of Texas in 2012, where she was awarded the Frank Gerth III Graduate Excellence Award. She was also awarded a J. William Fulbright scholarship between 2012 and 2014, and a University Graduate Continuing Fellowship for the academic year 2016/2017. In June 2017, Soledad will join the New York University Center for Data Science as a Research Fellow, and the Simons Collaboration Algorithms and Geometry as a Collaboration Scientist.

Permanent address: solevillar@gmail.com

This dissertation was typeset with LaTeXby the author.