Math. Program., Ser. A (2017) 165:605–642
DOI 10.1007/s10107-016-1097-0

FULL LENGTH PAPER

Probably certifiably correct k-means clustering

Takayuki Iguchi[1] · Dustin G. Mixon[1] ·
Jesse Peterson[1] · Soledad Villar[2]

Received: 23 April 2016 / Accepted: 23 November 2016 / Published online: 21 December 2016

Abstract Recently, Bandeira (C R Math, 2015) introduced a new type of algorithm (the so-called probably certifiably correct algorithm) that combines fast solvers with the optimality certificates provided by convex relaxations. In this paper, we devise such an algorithm for the problem of k-means clustering. First, we prove that Peng and Wei's semidefinite relaxation of k-means Peng and Wei (SIAM J Optim 18(1):186–205, 2007) is tight with high probability under a distribution of planted clusters called the stochastic ball model. Our proof follows from a new dual certificate for integral solutions of this semidefinite program. Next, we show how to test the optimality of a proposed k-means solution using this dual certificate in quasilinear time. Finally, we analyze a version of spectral clustering from Peng and Wei (SIAM J Optim 18(1):186–205, 2007) that is designed to solve k-means in the case of two clusters. In particular, we show that this quasilinear-time method typically recovers planted clusters under the stochastic ball model.

Mathematics Subject Classification 65-XX · 90-XX · 46N10 · 68Q87

1 Introduction

Clustering is a central problem in unsupervised machine learning. It consists of partitioning a given finite set $\{x_i\}_{i=1}^N$ of points in $\mathbb{R}^m$ into k subsets such that some dissimilarity function is minimized. Usually, this function is chosen ad hoc with an application in mind. A particularly common choice is the *k-means objective*:

✉ Dustin G. Mixon
dustin.mixon@gmail.com

[1] Department of Mathematics and Statistics, Air Force Institute of Technology, Wright-Patterson AFB, OH, USA

[2] Department of Mathematics, University of Texas at Austin, Austin, TX, USA

$$\begin{aligned} \text{minimize} \quad & \sum_{t=1}^{k} \sum_{i \in A_t} \left\| x_i - \frac{1}{|A_t|} \sum_{j \in A_t} x_j \right\|_2^2 \\ \text{subject to} \quad & A_1 \sqcup \cdots \sqcup A_k = \{1, \ldots, N\} \end{aligned} \tag{1}$$

Problem (1) is NP-hard in general [14]. A popular heuristic for solving k-means is Lloyd's algorithm, also known as the k-means algorithm [16]. This algorithm alternates between calculating centroids of proto-clusters and reassigning points according to the nearest centroid. In general, Lloyd's algorithm (and its variants [3,21]) may converge to local minima of the k-means objective (e.g., see Sect. 5 of [4]). Furthermore, the output of Lloyd's algorithm does not indicate how far it is from optimal. Instead, we seek a new sort of algorithm, recently introduced by Bandeira [5]:

Definition 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 (**P**, **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.

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. Along these lines, the k-means problem enjoys a semidefinite relaxation. To see this, let 1_A denote the indicator function of $A \subseteq \{1, \ldots, N\}$ (i.e. $1_A \in \{0, 1\}^N$ with $(1_A)_i = 1$ if and only if $i \in A$), and define the $N \times N$ matrix D by $D_{ij} := \|x_i - x_j\|_2^2$. Then taking

$$X := \sum_{t=1}^{k} \frac{1}{|A_t|} 1_{A_t} 1_{A_t}^\top, \tag{2}$$

the k-means objective (1) may be expressed as $\frac{1}{2}\operatorname{Tr}(DX)$. Since X satisfies several convex constraints, we may relax the region of optimization to produce a convex program, namely, the Peng–Wei semidefinite relaxation [22] (cf. [6]):

$$\begin{aligned} \text{minimize} \quad & \operatorname{Tr}(DX) \\ \text{subject to} \quad & \operatorname{Tr}(X) = k, \quad X1 = 1, \quad X \geq 0, \quad X \succeq 0 \end{aligned} \tag{3}$$

Here, $X \geq 0$ means that each entry of X is nonnegative, whereas $X \succeq 0$ means that X is symmetric and positive semidefinite. A similar relaxation for clustering with regularization appears in [25].

Recently, it was shown that under a certain random data model, this convex relaxation is *tight* with high probability [4], that is, every solution to the relaxed problem (3) has the form (2), thereby identifying an optimal clustering. As such, in this high-probability event, one may solve the dual 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 s, whereas a CVX implementation [11] of the dual of (3) for the same data takes about 20 s. 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 [20]), it is far too slow to wait for in practice.

As a fast alternative, Bandeira [5] recently devised a general technique to certify global optimality. This technique 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 to be suitable choices for (i) and (ii), respectively. For (iii), one might adapt Bandeira's original method in [5] based on complementary slackness (see Fig. 1 for an illustration). In this paper, we provide a theoretical basis for each of these components in the context of k-means.

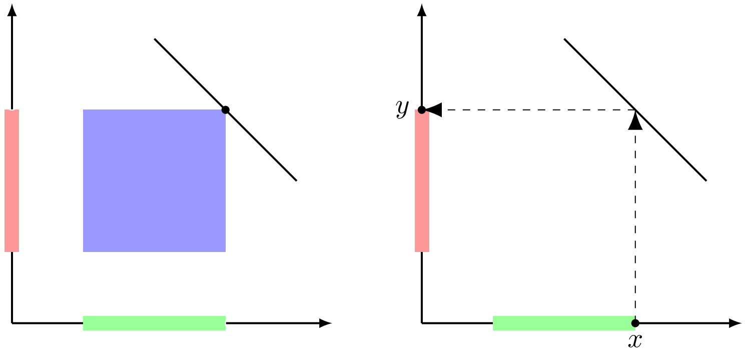

Fig. 1 (*left*) Depiction of complementary slackness. The *horizontal axis* represents a vector space in which we consider a cone program (e.g., a linear or semidefinite program), and the feasibility region of this program is highlighted in *green*. The dual program concerns another vector space, which 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 [5]. 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) (color figure online)

Table 1 Summary of cluster recovery guarantees under the stochastic ball model

Method	Sufficient condition	Optimal?	References
Thresholding	$\Delta > 4$	Yes	(Simple exercise)
k-Medians LP	$\Delta \geq 4$	No	Theorem 2 in [9]
	$\Delta \geq 3.75$	No	Theorem 1 in [19]
	$\Delta > 2$	Yes	Theorem 1 in [4]
k-Means LP	$\Delta > 4$	Yes	Theorem 9 in [4]
k-Means SDP	$\Delta > 2\sqrt{2}(1 + 1/\sqrt{m})$	No	Theorem 3 in [4]
	$\Delta > 2 + k^2/m$	No	Theorem 9
Spectral k-means	$\Delta > \Delta^\star, k = 2$	Yes	Theorem 14

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 $1 - e^{-\Omega_{\mathcal{D},\gamma}(n)}$

1.1 Technical background and overview

The first two components of a probably certifiably correct algorithm require nonconvex and convex solvers that perform well under some "reasonable" distribution of problem instances. In the context of geometric clustering, it has become popular recently to consider a particular model of data called the *stochastic ball model*, introduced in [19]:

Definition 2 ($(\mathcal{D}, \gamma, n)$-*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 $\mathcal{D}$ supported on the unit ball. The points from cluster a are then taken to be $x_{a,i} := r_{a,i} + \gamma_a$. We denote $\Delta := \min_{a \neq b} \|\gamma_a - \gamma_b\|_2$.

Table 1 summarizes the state of the art for recovery guarantees under the stochastic ball model. In [19], it was shown that an LP relaxation of k-medians will, with high probability, recover clusters drawn from the stochastic ball model provided the smallest distance between ball centers is $\Delta \geq 3.75$. Note 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. For the k-means problem, Awasthi et al. [4] studies the Peng–Wei semidefinite relaxation and demonstrates exact recovery in the regime $\Delta > 2\sqrt{2}(1 + 1/\sqrt{m})$, where m is the dimension of the Euclidean space.

As indicated in Table 1, we also study the Peng–Wei SDP, but our guarantee is different from [4]. In particular, we demonstrate tightness in the regime $\Delta > 2+k^2/m$, which is near-optimal for large m. The source of this improvement is a different choice of dual certificate, which leads to the following result (see Sect. 2 for details):

Theorem 3 *Take X of the form (2), 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 (3) *if*

$$P_{\Lambda^\perp} Z P_{\Lambda^\perp} \preceq z P_{\Lambda^\perp}. \tag{4}$$

To prove that $\Delta > 2 + k^2/m$ suffices for the SDP to recover the planted clustering under the stochastic ball model, we estimate the left- and right-hand sides of (4) with the help of standard techniques from random matrix theory and concentration of measure; see Appendix 3 for the (rather technical) details. While this is an improvement over the condition from [4] in the large-m regime, there are other regimes (e.g., $k = m$) for which their condition is much better, leaving open the question of what the optimal bound is. Conjecture 4 in [4] suggests that $\Delta > 2$ suffices for the k-means SDP to recover planted clusters under the stochastic ball model, but as we illustrate in Sect. 2.3, this conjecture is surprisingly false.

Having accomplished component (ii) in Bandeira's PCC technique, we tackle component (iii) next. For this, consider the matrix

$$A := \frac{z}{N} 11^\top + P_{\Lambda^\perp} Z P_{\Lambda^\perp}, \tag{5}$$

where z and Z come from Theorem 3. 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 3. As such, component (iii) can be accomplished by solving the following fundamental problem from linear algebra:

Problem 4 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 [5], but it was left unresolved. In this paper, 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 Sect. 3, 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) and Theorem 4 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). See Fig. 2 for an illustration of the runtime of our method. 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 have no idea how to accomplish this in linear or even near-linear time.

Now that we have discussed components (ii) and (iii) in Bandeira's PCC technique, we conclude by discussing component (i). While we presume that there exists a fast

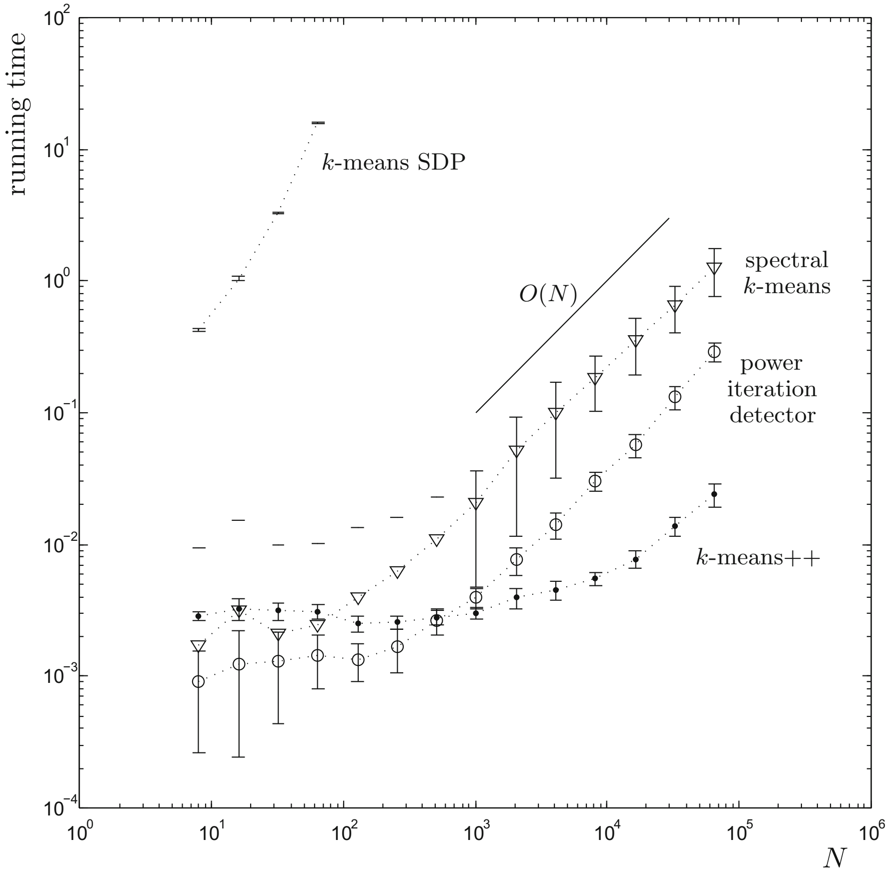

Fig. 2 Running time of different algorithms as a function of number of points to cluster. *Points* are sampled from two unit balls in $\mathbb{R}^6$ at distance 2.3 apart. For each $N \in \{2^3, 2^4, \dots, 2^{16}\}$, we perform 300 trials of the following experiment: draw $N/2$ points uniformly at random from each ball, and then compute four different functions: (a) MATLAB's built-in implementation of k-means++, (b) a CVX implementation [11] of the k-means SDP (3), (c) the power iteration detector (Algorithm 1) with A given by (5), and (d) spectral k-means clustering (Algorithm 2). For each trial, we recorded the runtime in seconds. Above, we plot the average runtime along with *vertical error bars* for standard deviation. For the record, the power iteration detector failed to certify optimality (i.e., reject H_0 in (16)) in at most 3% of the trials with $N \leq 2^7$, but provided a certificate of optimality in every trial otherwise; similarly, spectral k-means failed to recover the planted clusters in two of the trials with $N = 2^3$. In the rest of the trials, the planted clusters were recovered. Our implementation of the k-means SDP was too slow to perform trials with $N \geq 2^7$ in a reasonable amount of time, so we only recorded runtimes for $N \in \{2^3, 2^4, 2^5, 2^6\}$. As the plot illustrates, the other algorithms ran in quasilinear time, as expected

initialization of Lloyd's algorithm that performs well under the stochastic ball model, we leave this investigation for future research. Instead, Sect. 4 considers a spectral method introduced by Peng and Wei [22]. We show that when $k = 2$, this method performs as well as the optimizer of the original k-means problem under the stochastic ball model. Figure 2 illustrates the quasilinear runtime of this approach.

1.2 Outline

In this paper, we provide a theoretical analysis of probably certifiably correct k-means clustering, and we do so by developing components (i), (ii) and (iii) of Bandeira's general technique. First, we investigate (ii) in Sect. 2 by analyzing the tightness of the Peng–Wei SDP. In particular, we choose a different dual certificate from the one used in [4], and our choice demonstrates tightness in the SDP for clusters that are near-optimally close. Section 3 then addresses (iii) by providing a fast method of computing this dual certificate given the optimal k-means partition. In fact, a subroutine of our method (the so-called power iteration detector) resolves a gap in Bandeira's original PCC theory [5], and as such, we expect this to be leveraged in future PCC algorithms. We conclude in Sect. 4 with some theoretical guarantees for (i). Here, we focus on the case $k = 2$, and we show that a slight modification of the spectral clustering–based method in [22] manages to recover the optimal k-means partition with high probability under the stochastic ball model. We conclude in Sect. 5 with a discussion of various open problems.

2 A typically tight relaxation of k-means

This section establishes that the Peng–Wei semidefinite relaxation (3) of the k-means problem (1) is typically tight under the stochastic ball model. First, we find a deterministic condition on the set of points under which the relaxation finds the k-means-optimal solution. Later, we discuss when this deterministic condition is satisfied with high probability under the stochastic ball model.

2.1 The dual program

To derive the dual of (3), we will leverage the general setting from cone programming [18], namely, that given closed convex cones K and L, the dual of (6) is given by (7):

$$\begin{array}{ll} \text{maximize} & \langle c, x\rangle \\ \text{subject to} & b - Ax \in L \\ & \qquad x \in K \end{array} \tag{6}$$

$$\begin{array}{ll} \text{minimize} & \langle b, y\rangle \\ \text{subject to} & A^*y - c \in K^* \\ & \qquad\quad y \in L^* \end{array} \tag{7}$$

where A^* denotes the adjoint of A, while K^* and L^* denote the dual cones of K and L, respectively. In our case, $c = -D$, $x = X$, and K is simply the cone of positive semidefinite matrices (as is K^*). Before we determine L, we need to interpret the remaining constraints in (3). To this end, we note that $\mathrm{Tr}(X) = k$ is equivalent to $\langle X, I\rangle = k$, $X1 = 1$ is equivalent to having

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

and $X \geq 0$ is equivalent to having

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

(These last two equivalences exploit the fact that X is symmetric.) As such, we can express the remaining constraints in (3) using a linear operator 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 $\{\frac{1}{2}(e_i e_j^\top + e_j e_i^\top)\}_{i,j=1,i\leq j}^N$. The remaining constraints in (3) are equivalent to having $b - Ax \in L$, where $b = k \oplus 1 \oplus 0$ and $L = 0 \oplus 0 \oplus \mathbb{R}_{\geq 0}^{N(N+1)/2}$. Writing $y = z \oplus \alpha \oplus (-\beta)$, the dual of (3) is then given by

$$\begin{aligned} \text{minimize} \quad & kz + \sum_{i=1}^N \alpha_i \\ \text{subject to} \quad & Q := zI + \sum_{i=1}^N \alpha_i \cdot \tfrac{1}{2}\left(e_i 1^\top + 1 e_i^\top\right) \\ & -\sum_{i=1}^N \sum_{j=i}^N \beta_{ij} \cdot \tfrac{1}{2}\left(e_i e_j^\top + e_j e_i^\top\right) + D \succeq 0 \\ & \beta \geq 0 \end{aligned} \tag{8}$$

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

$$\beta := \sum_{i=1}^N \sum_{j=i}^N \beta_{ij} \cdot \frac{1}{2}\left(e_i e_j^\top + e_j e_i^\top\right).$$

Indeed, when $i \leq j$, 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 (whose proof has been reproduced in Appendix 1 for completeness) characterizes all possible dual certificates of (3):

Proposition 5 (Theorem 4 in [13], cf. [4]) *Take* $X := \sum_{a=1}^k \frac{1}{n_a} 1_a 1_a^\top$, *where* n_a *denotes the number of points in cluster* a. *The following are equivalent:*

(a) *X is a solution to the semidefinite relaxation* (3).

(b) *Every solution to the dual program* (8) *satisfies*

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

(c) *Every solution to the dual program* (8) *satisfies*

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

Remark 1 (Pointed out by Xiaodong Li on an earlier version of this manuscript) 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 $R^{(a,a)} = 1_a 1_a^T$, $R^{(b,b)} = I_b$, $R^{(b,a)} = 0$ for all $b \neq 0$. Then $L := R^\top QR$ 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 $L^{(b,a)} = Q^{(b,a)}1_a^\top 1_a$. □

The following subsection will leverage Proposition 5 to identify a condition on D that implies that the SDP (3) relaxation is tight.

2.2 Selecting a dual certificate

The goal is to certify when the SDP relaxation is tight. In this event, Proposition 5 characterizes acceptable dual certificates (z,α,β), but this information fails to uniquely determine a certificate. 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 5. 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

$$\begin{aligned}
&\left(\sum_{t=1}^{k}\sum_{i\in t}\alpha_{t,i}\cdot\frac{1}{2}(e_{t,i}1^\top + 1e_{t,i}^\top)\right)^{(a,b)}\\
&\quad= \sum_{i\in a}\alpha_{a,i}\cdot\frac{1}{2}e_i1^\top + \sum_{j\in b}\alpha_{b,j}\cdot\frac{1}{2}1e_j^\top\\
&\quad= -\frac{1}{2}\left(\frac{1}{n_a}+\frac{1}{n_b}\right)z + \sum_{i\in a}\left(\frac{1}{n_a^2}1^\top D^{(a,a)}1 - \frac{2}{n_a}e_i^\top D^{(a,a)}1\right)\frac{1}{2}e_i1^\top\\
&\qquad + \sum_{j\in b}\left(\frac{1}{n_b^2}1^\top D^{(b,b)}1 - \frac{2}{n_b}e_j^\top D^{(b,b)}1\right)\frac{1}{2}1e_j^\top,
\end{aligned}$$

and so since

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

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

$$\begin{aligned} E^{(a,b)} &:= \frac{1}{2}\left(\frac{1}{n_a} + \frac{1}{n_b}\right)11^\top \\ M^{(a,b)} &:= D^{(a,b)} + \sum_{i \in a}\left(\frac{1}{n_a^2}1^\top D^{(a,a)}1 - \frac{2}{n_a}e_i^\top D^{(a,a)}1\right)\frac{1}{2}e_i1^\top \qquad (9) \\ &\quad + \sum_{j \in b}\left(\frac{1}{n_b^2}1^\top D^{(b,b)}1 - \frac{2}{n_b}e_j^\top D^{(b,b)}1\right)\frac{1}{2}1e_j^\top \\ B^{(a,b)} &= \frac{1}{2}\beta^{(a,b)} \qquad (10) \end{aligned}$$

for every $a, b \in \{1, \ldots, k\}$. 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 $a \in \{1, \ldots, k\}$. 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 6 *Let E be the matrix defined by* (9)*. Then* $\operatorname{rank}(E) \in \{1, 2\}$*. The eigenvalue of largest magnitude is $\lambda \geq k$, 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 $\operatorname{rank}(E) \in \{1, 2\}$, and it is easy to calculate $1^\top E1 = Nk$ and $\operatorname{Tr}(E) = k$. Observe that

$$\lambda = \sup_{\substack{x \in \mathbb{R}^N \\ \|x\|_2 = 1}} x^\top Ex \geq \frac{1}{N}1^\top E1 = k,$$

and combining with $\operatorname{rank}(E) \leq 2$ and $\operatorname{Tr}(E) = k$ 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 $\operatorname{span}\{1_a\}_{a=1}^k$ 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 denote the corresponding eigenvector. Then combining Lemma 6 and Remark 1 we know v is also an eigenvector of $M - B$.

Since

$$0 = (Q1_b)_a = \Big((z(I - E) + M - B)1_b\Big)_a$$
$$= -zE^{(a,b)}1 + M^{(a,b)}1 - B^{(a,b)}1 = -z\frac{n_a + n_b}{2n_a}1 + M^{(a,b)}1 - B^{(a,b)}1, \tag{11}$$

then, in order for there to exist a vector $B^{(a,b)}1 \geq 0$ that satisfies (11), z must satisfy

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

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

$$z \leq \min_{\substack{a,b\in\{1,\ldots,k\}\\ a\neq b}} \frac{2n_a}{n_a + n_b}\min(M^{(a,b)}1). \tag{12}$$

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 (11) for every (a, b), even when z satisfies equality in (12). Indeed, we define

$$u_{(a,b)} := M^{(a,b)}1 - z\frac{n_a + n_b}{2n_a}1, \qquad \rho_{(a,b)} := u_{(a,b)}^\top 1, \qquad B^{(a,b)} := \frac{1}{\rho_{(b,a)}}u_{(a,b)}u_{(b,a)}^\top \tag{13}$$

for every $a, b \in \{1, \ldots, k\}$ with $a \neq b$. Then by design, B immediately satisfies (11). 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 $u_{(a,b)} \geq 0$ (and thus $\rho_{(a,b)} \geq 0$) by (12), and we implicitly require $\rho_{(a,b)} > 0$ for division to be permissible. As such, we also have $B^{(a,b)} \geq 0$, 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

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

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

Theorem 7 *Take $X := \sum_{t=1}^k \frac{1}{n_t}1_t1_t^\top$, where n_t denotes the number of points in cluster t. Consider M defined by (10), pick z so as to satisfy equality in (12), take B defined by (13), and let Λ denote the span of $\{1_t\}_{t=1}^k$. Then X is a solution to the semidefinite relaxation (3) if*

$$P_{\Lambda^\perp}(B - M)P_{\Lambda^\perp} \preceq zP_{\Lambda^\perp}. \tag{14}$$

The next subsection leverages this sufficient condition to establish that the Peng–Wei SDP (3) is typically tight under the stochastic ball model.

2.3 Integrality of the relaxation under the stochastic ball model

We first note that our sufficient condition (14) is implied by

$$\|P_{\Lambda^\perp} M P_{\Lambda^\perp}\| + \|P_{\Lambda^\perp} B P_{\Lambda^\perp}\| \leq z$$

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 Appendix 2), we arrive at the following corollary:

Corollary 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 (10), pick z so as to satisfy equality in (12), and take $\rho_{(a,b)}$ defined by (13). Then X is a solution to the semidefinite relaxation (3) if

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

In Appendix 3, we leverage the stochastic ball model to bound each term in Corollary 8, and in doing so, we identify a regime in which the data points typically satisfy the sufficient condition given in Corollary 8:

Theorem 9 *The k-means semidefinite relaxation (3) recovers the planted clusters in the $(\mathcal{D}, \gamma, n)$-stochastic ball model with probability $1 - e^{-\Omega_{\mathcal{D},\gamma}(n)}$ provided $\Delta > 2 + k^2/m$.*

We note that Theorem 9 is an improvement to the main result of the authors' preprint [13]. When $k = o(m^{1/2})$, Theorem 9 is near-optimal, and in this sense, it's a significant improvement over the sufficient condition

$$\Delta > 2\sqrt{2}\left(1 + \frac{1}{\sqrt{m}}\right) \tag{15}$$

given in [4]. However, there are regimes (e.g., $k = m$) for which (15) is much better, leaving open the question of what the optimal bound is. Conjecture 4 in [4] suggests that $\Delta > 2$ suffices for the k-means SDP to recover planted clusters under the stochastic ball model, but as we illustrate below, this conjecture is surprisingly false.

Consider the special case where $m = 1$, $\mathcal{D}$ is uniform on $\{\pm 1\}$, and $k = 2$. Centering the two balls on $\pm\Delta/2$, then all of the points land in $\{\pm\Delta/2 \pm 1\}$. The k-means-optimal clustering will partition the real line into two semi-infinite intervals, and so there are three possible ways of clustering these points. Suppose exactly $N/4$ of the points land in each of the 4 positions. Then by symmetry, there are only two ways to cluster:

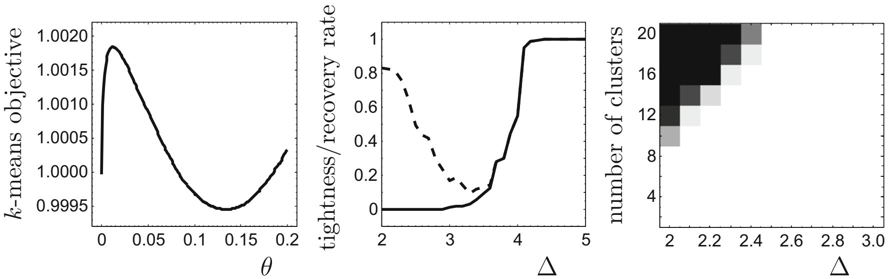

Fig. 3 (*left*) Take two unit disks in $\mathbb{R}^2$ with centers on the x-axis at distance 2.08 apart. Let x_0 denote the smallest possible x-coordinate in the disk on the right. For each disk, draw $N/2 = 50{,}000$ points uniformly at random from the perimeter. Given θ, cluster the points according to whether the x-coordinate is smaller than $x_0 + \theta$. When $\theta = 0$, this clustering gives the planted clusters, and the k-means objective (divided by N) is 1. We plot this normalized k-means objective for $\theta \in [0, 0.2]$. Since N is large, this curve is very close to its expected shape, and we see that there are clusters whose k-means value is smaller than that of the planted clustering. (*center*) Take two intervals of width 2 in $\mathbb{R}$, and let Δ denote the distance between the midpoints of these intervals. For each interval, draw 10 points at random from its endpoints, and then run the k-means SDP. For each $\Delta = 2 : 0.1 : 5$, after running 2000 trials of this experiment, we plot the proportion of trials for which the SDP relaxation was tight (*dashed line*) and the proportion of trials for which the SDP recovered the planted clusters (*solid line*). In this case, cluster recovery appears to exhibit a phase transition at $\Delta = 4$. (*right*) For each $\Delta = 2 : 0.1 : 3$ and $k = 2 : 2 : 20$, consider the unit balls in $\mathbb{R}^{20}$ centered at $\{\frac{\Delta}{\sqrt{2}} e_i\}_{i=1}^k$, where e_i denotes the ith identity basis element. Draw 100 points uniformly from each ball, and test if the resulting data points satisfy (14). After performing 10 trials of this experiment for each (Δ, k), we shade the corresponding pixel according to the proportion of successful trials (white means every trial satisfied (14)). This *plot* indicates that our certificate (14) is to blame for Theorem 9's dependence on k

either we select the planted clusters, or we make the left-most location its own cluster. Interestingly, a little algebra reveals that this second alternative is strictly better in the k-means sense provided $\Delta < 1 + \sqrt{3} \approx 2.7320$. Also, in this regime, then as N gets large, the proportion of points in each position will be so close to $1/4$ (with high probability) that this clustering will beat the planted clusters.

Overall, when $m = 1$ and $k = 2$, then $\Delta \geq 1 + \sqrt{3}$ is necessary for minimizing the k-means objective to recover planted clusters for an arbitrary $\mathcal{D}$. As a relaxation, the k-means SDP recovers planted clusters only if minimizing the k-means objective does so as well, and so it inherits this necessary condition, thereby disproving Conjecture 4 in [4]. Furthermore, as Fig. 3(left) illustrates, a similar counterexample is available in higher dimensions.

To study when the SDP recovers the clusters, let's continue with the case where $m = 1$ and $k = 2$. We know that minimizing k-means will recover the clusters with high probability provided $\Delta > 1 + \sqrt{3}$. However, Theorem 9 only guarantees that the SDP recovers the clusters when $\Delta > 6$; in fact, (15) is slightly better here, giving that $\Delta \geq 5.6569$ suffices. To shed light on the disparity, Fig. 3(center) illustrates the performance of the SDP for different values of Δ. Observe that the SDP is often tight when Δ is close to 2, but it doesn't reliably recover the planted clusters until $\Delta > 4$. We suspect that $\Delta = 4$ is a phase transition for cluster recovery in this case.

Qualitatively, the biggest difference between Theorem 9 and (15) is the dependence on k that Theorem 9 exhibits. Figure 3(right) illustrates that this comes from (14), meaning that one would need to use a completely different dual certificate in order to remove this dependence.

3 A fast test for *k*-means optimality

In this section, we leverage the certificate (14) 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).

3.1 Leading eigenvector hypothesis test

This subsection concerns Problem 4. To solve this problem, one might be inclined to apply the power method:

Proposition 10 (Theorem 8.2.1 in [10]) *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with eigenvalues $\{\lambda_i\}_{i=1}^n$ (counting multiplicities) satisfying*

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

and with corresponding orthonormal eigenvectors $\{v_i\}_{i=1}^n$. Pick a unit-norm vector $q_0 \in \mathbb{R}^n$ and consider the power iteration $q_{j+1} := Aq_j/\|Aq_j\|_2$. If q_0 is not orthogonal to v_1, then

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

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

$$\begin{aligned} &H_0: \text{span}(v) \text{ is not the unique leading eigenspace of } A,\\ &H_1: \text{span}(v) \text{ is the unique leading eigenspace of } A. \end{aligned} \tag{16}$$

To test these hypotheses, pick a tolerance $\epsilon > 0$ and run the power iteration detector (see Algorithm 1). This detector terminates either by accepting H_0 or by rejecting H_0

Algorithm 1: 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 (16)
$\lambda \leftarrow v^\top Av$
Draw $q$ uniformly at random from the unit sphere in $\mathbb{R}^n$
while no decision has been made do
    if $|q^\top Aq| > |\lambda|$ then
        Print accept $H_0$
    else if $(v^\top q)^2 \geq 1-\epsilon$ then
        Print reject $H_0$ and accept $H_1$
    end
    $q \leftarrow Aq/\|Aq\|_2$
end
```

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:

Definition 11 Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$ and unit eigenvector v of A, put $\lambda = v^\top Av$, 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 ≥ 2,
(b) $-\lambda$ is an eigenvalue of A, or
(c) $-\lambda_1$ is an eigenvalue of A.

Theorem 12 *Consider the power iteration detector (Algorithm 1), let q_j denote q at the jth 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:*

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

(iii) *If H_1 holds, then*

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

with probability $\geq 1-\pi_\epsilon$, 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 $|\lambda_1| \geq \cdots \geq |\lambda_n|$, 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 10 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 [10], 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

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

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

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

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

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

As such, $\epsilon > (v_1^\top q_j)^2 \geq (v_1^\top q_0)^2$. 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 10 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 10, it suffices to have

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

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

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

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

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

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 13 *If $\epsilon \geq n^{-1}e^{-2n}$, then $\pi_\epsilon \leq 3\sqrt{n\epsilon}$.*

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

$$\Pr\Big(Z^2 < a\Big) \leq \sqrt{\frac{2a}{\pi}}.$$

Also, Lemma 1 in [15] gives

$$\Pr\Big(Q \geq n + 2\sqrt{nx} + 2x\Big) \leq e^{-x} \qquad \forall x > 0.$$

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

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

□

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 $|\lambda_2| \leq (1-\delta)|\lambda_1|$.

3.2 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 (14) holds if and only if $v := \frac{1}{\sqrt{N}}1$ is a leading eigenvector of the matrix

$$A := \frac{z}{N}11^\top + P_{\Lambda^\perp}(B - M)P_{\Lambda^\perp} = \frac{z}{N}11^\top + P_{\Lambda^\perp}(B - D)P_{\Lambda^\perp}. \tag{17}$$

(The second equality follows from distributing the $P_{\Lambda^\perp}$'s and recalling the definition of M in (10).) As such, it suffices that (A, v) satisfy H_1 in (16). 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 (14). In particular, a positive test with tolerance ϵ will yield $\geq 1 - \pi_\epsilon$ 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 (17):

Steps in computation of $x \mapsto Ax$.	*cost in operations*
1: Compute $\nu_a \in \mathbb{R}^{n_a}$ such that $(\nu_a)_i = \lVert x_{a,i}\rVert_2^2$ 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 $\mu_a = \frac{1}{2}(\frac{1}{n_a^2}11^\top - \frac{2}{n_a}I)D^{(a,a)}1$ 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, and as we illustrate in Fig. 2, the runtimes are comparable.)

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* [5]. 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 [1].

Given A, put $B := 2A - 11^\top + I$, 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 [5], Bandeira observes that, given a partition matrix X by some means (such as the fast algorithm provided in [2]), 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 [1]. 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

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

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.

4 A fast *k*-means solver for two clusters

The previous section illustrated how to quickly test whether a proposed solution to the k-means problem is optimal. In particular, this test will be successful with high probability if the data follows the stochastic ball model with $\Delta > 2+k^2/m$. It remains to find a fast k-means solver which also performs in this regime.

In doing so, we maintain the philosophy that our algorithm should not "see" the stochastic ball model. Indeed, we view the stochastic ball model as a method of evaluating clustering algorithms rather than a realistic data model. For example, Lloyd's algorithm can be viewed as an alternating minimization of the lifted objective function:

$$f(A_1, \ldots, A_k, c_1, \ldots, c_k) := \sum_{t=1}^{k} \sum_{i \in A_t} \|x_i - c_t\|^2,$$
$$A_1 \sqcup \cdots \sqcup A_k = \{1, \ldots, N\},\ c_1, \ldots, c_k \in \mathbb{R}^m,$$

and since this function is minimized at the k-means optimizer (regardless of how the data is distributed), such an algorithm is acceptable. On the other hand, one might consider matching the stochastic ball model to the data by maximizing the following function:

$$g(c_1, \ldots, c_k) := \sum_{i=1}^{N} \sum_{t=1}^{k} p_{\mathcal{D}}(x_i - c_t), \qquad c_1, \ldots, c_k \in \mathbb{R}^m,$$

where $p_{\mathcal{D}}(\cdot)$ denotes the density function of $\mathcal{D}$, which is supported on the unit ball centered at the origin. One could certainly devise a fast greedy method such as matching pursuit [17] to optimize this objective function (especially if $p_{\mathcal{D}}$ is smooth), but doing so violates our philosophy.

In [22], Peng and Wei showed that k-means is equivalent to the following program:

$$\begin{array}{ll} \text{minimize} & \operatorname{Tr}(DX) \\ \text{subject to} & X^\top \ = X,\ X^2 = X,\ \operatorname{Tr}(X) = k,\ X1 = 1,\ X \geq 0 \end{array} \tag{18}$$

One may quickly observe that the SDP (3) we analyzed in Sect. 2 is a relaxation of this program. In this section, we follow Peng and Wei [22] by considering another relaxation of (18), obtained by discarding the $X \geq 0$ constraint (this is known as the *spectral clustering* relaxation [7,8]). We first denote the $m \times N$ matrix $\Phi = [x_1 \cdots x_N]$. Without loss of generality, the data set is centered at the origin so that $\Phi 1 = 0$. Letting ν denote the $N \times 1$ vector with $\nu_i = \|x_i\|_2^2$, then

$$D_{ij} = \|x_i - x_j\|_2^2 = \|x_i\|_2^2 - 2x_i^\top x_j + \|x_j\|_2^2 = \left(\nu 1^\top - 2\Phi^\top \Phi + 1\nu^\top\right)_{ij}.$$

As such, $D = \nu 1^\top - 2\Phi^\top \Phi + 1\nu^\top$, and so the constraints $X = X^\top$ and $X1 = 1$ together imply an alternative expression for the objective function:

$$\begin{aligned}\operatorname{Tr}(DX) &= \operatorname{Tr}\left(\nu 1^\top X - 2\Phi^\top \Phi X + 1\nu^\top X\right)\\ &= \operatorname{Tr}\left(\nu 1^\top X^\top\right) - 2\operatorname{Tr}\left(\Phi^\top \Phi X\right) + \operatorname{Tr}\left(X1\nu^\top\right)\\ &= 2\nu^\top 1 - 2\operatorname{Tr}\left(\Phi^\top\Phi X\right).\end{aligned}$$

We conclude that minimizing $\operatorname{Tr}(DX)$ is equivalent to maximizing $\operatorname{Tr}(\Phi^\top \Phi X)$.

Next, we observe that the feasible X in our relaxation are precisely the rank-k $N \times N$ orthogonal projection matrices satisfying $X1 = 1$. This in turn is equivalent to X having the form $X = \frac{1}{N}11^\top + Y$, where Y is a rank-$(k-1)$ $N \times N$ orthogonal projection matrix satisfying $Y1 = 0$. Discarding the $Y1 = 0$ constraint produces the following relaxation of (18):

$$\begin{aligned}&\text{maximize} \quad \operatorname{Tr}\left(\Phi^\top \Phi Y\right)\\ &\text{subject to} \quad Y^\top = Y, \quad Y^2 = Y, \quad \operatorname{Tr}(Y) = k-1\end{aligned} \tag{19}$$

For general values of k, this program amounts to finding $k-1$ principal components of the data. Recalling our initial clustering goal, after finding the optimal Y, it remains to take $X = \frac{1}{N}11^\top + Y$ and then round to a nearby member of the feasibility region in (18). In [22], Peng and Wei focus on the $k = 2$ case; they reduce the rounding step to a 2-means problem on the real line, and they establish an approximation ratio of 2 for this relax-and-round procedure. Here, we are concerned with exact recovery under the stochastic ball model, and as such, we slightly modify the rounding step.

When $k = 2$, the solution to (19) has the form $Y = yy^\top$, where y is a leading unit eigenvector of $\Phi^\top\Phi$. Our task is to find a matrix of the form $\frac{1}{|A|}1_A 1_A^\top + \frac{1}{|B|}1_B 1_B^\top$ with $A \sqcup B = \{1, \ldots, N\}$ that is close to $\frac{1}{N}11^\top + yy^\top$. To this end, it seems natural to consider

$$A_\theta := \{i : y_i < \theta\}, \qquad B_\theta := A_\theta^c$$

for some threshold θ. Since the data is centered ($\Phi 1 = 0$), one may be inclined to take $\theta = 0$, but this will be a poor choice if the true clusters have significantly different numbers of points. Instead, we select the θ which minimizes the k-means objective of (A_θ, B_θ). Since we only need to consider $N-1$ choices of θ, this is plausibly tractable, although computing the k-means objective once costs $O(mN)$ operations, and so some care is necessary to keep the algorithm fast.

We will show how to find the optimal (A_θ, B_θ) in $O((m + \log N)N)$ operations using a simple dynamic program. Order the indices so that $y_1 \le \cdots \le y_N$. Then the function to minimize is

$$f(i) := \frac{1}{i}\underbrace{\sum_{j=1}^{i}\sum_{j'=1}^{i}\|x_j - x_{j'}\|_2^2}_{v_i} + \frac{1}{N-i}\underbrace{\sum_{j=i+1}^{N}\sum_{j'=i+1}^{N}\|x_j - x_{j'}\|_2^2}_{v_i^c}.$$

Algorithm 2: Spectral k-means clustering (for two clusters)

```
Input: $m \times N$ matrix $\Phi = [x_1 \cdots x_N]$ of points to be clustered
Output: Clusters $A \sqcup B = \{1, \ldots, N\}$
Subtract centroid $\frac{1}{N}\sum_{i=1}^{N} x_i$ from each column of $\Phi$ to produce $\Phi_0$
Compute leading eigenvector $y$ of $\Phi_0^\top \Phi_0$
Find $\theta$ that minimizes the $k$-means objective of $(\{i : y_i < \theta\}, \{i : y_i \geq \theta\})$
$(A, B) \leftarrow (\{i : y_i < \theta\}, \{i : y_i \geq \theta\})$
```

Expanding the square and distributing sums gives

$$v_{i+1} = v_i + 2\sum_{j=1}^{i} \|x_j\|_2^2 - 4x_{i+1}^\top \sum_{j=1}^{i} x_j + 2i\|x_{i+1}\|_2^2,$$

and the v_i^c's satisfy a similar recursion rule. As such, one may iteratively compute the v_i's and v_i^c's before computing the $f(i)$'s and then minimizing. Overall, the following procedure finds the optimal (A_θ, B_θ) in $O((m + \log N)N)$ operations:

```
STEPS                                                                     cost in operations
1: Sort the entries $y_1 \leq \cdots \leq y_N$.                                   $O(N \log N)$
2: Iteratively compute for every $i \in \{1, \ldots, N-1\}$:
   $s_1(i) := \sum_{j=1}^{i} x_j, \quad s_1^c(i) := \sum_{j=i+1}^{N} x_j, \quad s_2(i) := \sum_{j=1}^{i} \|x_j\|_2^2,$
   $s_2^c(i) := \sum_{j=i+1}^{N} \|x_j\|_2^2.$                                  $O(mN)$
3: Compute $v_1 = 0$ and $v_{i+1} = v_i + 2s_2(i) - 4x_{i+1}^\top s_1(i) + 2i\|x_{i+1}\|_2^2$
   for every $i \in \{1, \ldots, N-2\}$.                                        $O(mN)$
4: Compute $v_{N-1}^c = 0$ and $v_{i-1}^c = v_i^c + 2s_2^c(i) - 4x_i^\top s_1^c(i) + 2(N-i)\|x_i\|_2^2$
   for every $i \in \{N-1, \ldots, 2\}$.                                        $O(mN)$
5: Compute $f(i) = v_i/i + v_i^c/(N-i)$ for every $i \in \{1, \ldots, N-1\}$.      $O(N)$
6: Find $i$ that minimizes $f(i)$ and output $\{1, \ldots, i\}$ and $\{i+1, \ldots, N\}$.   $O(N)$
```

Note that in the special case where $m = 1$, the above method exactly solves the k-means problem when $k = 2$ in only $O(N \log N)$ operations, recovering the rounding step of Peng and Wei [22]. For comparison, [26] leverages more sophisticated dynamic programming for the $m = 1$ case, but k is arbitrary and the algorithm costs $O(kN^2)$ operations.

See Algorithm 2 for a summary of our relax-and-round procedure. As a spectral method, this algorithm enjoys quasilinear computational complexity; see Fig. 2 for an illustration. In particular, when computing the leading eigenvector of $\Phi_0^\top \Phi_0$, each matrix–vector multiply in the power method costs only $O(mN)$ operations. Furthermore, as the following result guarantees, this algorithm performs well under the stochastic ball model:

Theorem 14 *Let $\Delta^\star = \Delta^\star(\mathcal{D}, k)$ denote 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 $1 - e^{-\Omega_{\mathcal{D},\gamma}(n)}$. When $k = 2$, spectral k-means* *clustering (Algorithm 2) recovers planted clusters under the stochastic ball model with probability* $1 - e^{-\Omega_{\mathcal{D},\gamma}(n)}$ *provided* $\Delta > \Delta^\star$.

See Appendix 4 for the proof. The main idea is that the leading eigenvector of $\Phi_0\Phi_0^\top$ is biased towards the difference between the ball centers, and as the following lemma establishes, this bias encourages spectral k-means clustering to separate the planted clusters:

Lemma 15 *Take two clusters contained in unit balls centered at* γ *and* $-\gamma$ *with* $\|\gamma\|_2 > 1$. *If minimizing the k-means objective recovers these clusters, then spectral k-means clustering (Algorithm 2) also recovers them, provided the leading eigenvector* z *of* $\Phi_0\Phi_0^\top$ *satisfies* $|\gamma^\top z| > \|z\|_2$.

Proof Write $\Phi_0 = \Phi - \mu 1^\top$, put $\theta := -\mu^\top z$, and observe that $y = \Phi_0^\top z$ is a leading eigenvector of $\Phi_0^\top\Phi_0$. Then

$$y_i = (x_i - \mu)^\top z = x_i^\top z + \theta \tag{20}$$

for every i. Next, if $|\gamma^\top z| > \|z\|_2$, then a simple trigonometric argument gives that the balls (and therefore the planted clusters) are separated by the hyperplane orthogonal to z. Combined with (20), we then have that the clusters can be identified according to whether $y_i < \theta$ or $y_i > \theta$. It therefore suffices to minimize the k-means objective subject to partitions of this form (for arbitrary thresholds θ), as so spectral k-means clustering succeeds. □

5 Discussion

This paper discussed various facets of probably certifiably correct algorithms for k-means clustering. There are still many questions that have yet to be answered:

- Let $\Delta^\star(\mathcal{D}, k)$ denote 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 $1 - e^{-\Omega_{\mathcal{D},\gamma}(n)}$. What is $\Delta^\star$? It was conjectured in [4] that $\Delta^\star = 2$, but as we demonstrated in Sect. 2.3, this is not the case.
- Let $\Delta^\star_{\mathrm{SDP}}(\mathcal{D}, k)$ denote the smallest value for which $\Delta > \Delta^\star_{\mathrm{SDP}}$ implies that solving the k-means SDP recovers planted clusters under the $(\mathcal{D}, \gamma, n)$-stochastic ball model with probability $1 - e^{-\Omega_{\mathcal{D},\gamma}(n)}$. What is $\Delta^\star_{\mathrm{SDP}}$? Considering Sect. 2.3 and Fig. 3(center), we suspect the SDP exhibits a performance gap: $\Delta^\star_{\mathrm{SDP}} > \Delta^\star$.
- Is there a single dual certificate for the k-means SDP that typically certifies planted clusters under the stochastic ball model whenever $\Delta > \Delta^\star_{\mathrm{SDP}}$? Does this certification have a quasilinear-time implementation similar to Sect. 3.2?
- Is there a quasilinear-time k-means solver that typically solves k-means under the stochastic ball model whenever $\Delta > \Delta^\star$? In particular, is there a quasilinear-time initialization of Lloyd's algorithm that meets this specification? Following the philosophy of Sect. 4, such algorithms should be designed so as to not "see" the stochastic ball model.

Acknowledgements The authors thank the anonymous referees, whose suggestions significantly improved this paper's presentation and literature review. The authors also thank Afonso S. Bandeira and Nicolas Boumal for interesting discussions and valuable comments on an earlier version of this manuscript, and Xiaodong Li and Yang Li for interesting comments on our dual certificate. DGM was supported by an AFOSR Young Investigator Research Program award, NSF Grant No. DMS-1321779, and AFOSR Grant No. F4FGA05076J002. SV was supported by Rachel Ward's NSF CAREER award and AFOSR Young Investigator Research Program award. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

Appendix 1: Proof of Proposition 5

Proof (a) $\Leftrightarrow$ (b): By complementary slackness, (a) is equivalent to having both

$$\langle A^* y - c, X\rangle = 0 \tag{21}$$

and

$$\langle y, b - A(X)\rangle = 0. \tag{22}$$

Since $Q \succeq 0$, we have

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

with equality if and only if $Q1_a = 0$ for every $a \in \{1, \dots, k\}$. Next, we recall that $y = z \oplus \alpha \oplus (-\beta)$, $b - A(X) \in L = 0 \oplus 0 \oplus \mathbb{R}_{\geq 0}^{N(N+1)/2}$, and $b = k \oplus 1 \oplus 0$. As such, (22) is equivalent to β having disjoint support with $\{\langle X, \frac{1}{2}(e_i e_j^\top + e_j e_i^\top)\rangle\}_{i,j=1,i\leq j}^{N}$, i.e., $\beta^{(a,a)} = 0$ for every cluster a.

(b) $\Rightarrow$ (c): Take any solution to the dual SDP (8), and note that

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

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:

$$\begin{aligned} 0 &= e_r^\top Q^{(a,a)} 1 \\ &= e_r^\top \left(zI + \sum_{i\in a} \alpha_{a,i} \cdot \frac{1}{2}(e_i 1^\top + 1 e_i^\top) + D^{(a,a)} \right) 1 \\ &= z + \sum_{i\in a} \alpha_{a,i} \cdot \frac{1}{2}(e_r^\top e_i 1^\top 1 + e_r^\top 1 e_i^\top 1) + e_r^\top D^{(a,a)} 1 \\ &= z + \sum_{i\in a} \alpha_{a,i} \cdot \frac{1}{2}(n_a \delta_{ir} + 1) + e_r^\top D^{(a,a)} 1. \end{aligned} \tag{23}$$

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

$$\begin{aligned}
0 &= 1^\top Q^{(a,a)}1\\
&= 1^\top\Big(zI+\sum_{i\in a}\alpha_{a,i}\cdot\frac{1}{2}(e_i1^\top+1e_i^\top)+D^{(a,a)}\Big)1\\
&= n_az+\sum_{i\in a}\alpha_{a,i}\cdot\frac{1}{2}(1^\top e_i1^\top 1+1^\top 1e_i^\top 1)+1^\top D^{(a,a)}1\\
&= n_az+n_a\sum_{i\in a}\alpha_{a,i}+1^\top D^{(a,a)}1,
\end{aligned}$$

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 (23):

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

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

(c) ⇒ (a): Take any solution to the dual SDP (8). Then by assumption, the dual objective at this point is given by

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

i.e., the primal objective (3) evaluated at X. Since X is feasible in the primal SDP, we conclude that X is optimal by strong duality. □

Appendix 2: Proof of Corollary 8

It suffices to have

$$\|P_{\Lambda^\perp} M P_{\Lambda^\perp}\| + \|P_{\Lambda^\perp} B P_{\Lambda^\perp}\| \leq z. \tag{24}$$

We will bound the terms in (24) separately and then combine the bounds to derive a sufficient condition for Theorem 7. To bound the first term in (24), let ν be the $N \times 1$ vector whose (a,i)th entry is $\|x_{a,i}\|_2^2$, and let Φ be the $m \times N$ matrix whose (a,i)th column is $x_{a,i}$. Then

$$\begin{aligned} D_{(a,i),(b,j)} &= \|x_{a,i} - x_{b,j}\|_2^2 = \|x_{a,i}\|_2^2 - 2x_{a,i}^\top x_{b,j} + \|x_{b,j}\|_2^2 \\ &= (\nu 1^\top - 2\Phi^\top \Phi + 1\nu^\top)_{(a,i),(b,j)}, \end{aligned}$$

meaning $D = \nu 1^\top - 2\Phi^\top\Phi + 1\nu^\top$. With this, we appeal to the blockwise definition of M (10):

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

For the second term in (24), 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

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

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

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

At this point, we use the definition of B (13) 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 $u_{(a,b)}$ (13) and combining these estimates then produces the result.

Appendix 3: Proof Theorem 9

In this section, we apply the certificate from Corollary 8 to the $(\mathcal{D},\gamma,n)$-stochastic ball model (see Definition 2) to prove our main result. We will prove Theorem 9 with the help of several lemmas.

Lemma 16 *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:

$$\|c_a-\gamma_a\|_2<\epsilon\qquad w.p.\qquad 1-e^{-\Omega_{m,\epsilon}(n)}\tag{25}$$

$$\left|\frac{1}{n}\sum_{i=1}^n\|r_{a,i}\|_2^2-\mathbb{E}\|r\|_2^2\right|<\epsilon\qquad w.p.\qquad 1-e^{-\Omega_\epsilon(n)}\tag{26}$$

$$\left|\frac{1}{n}\sum_{i=1}^n\|x_{a,i}-O_{ab}\|_2^2-\mathbb{E}\|r+\gamma_a-O_{ab}\|_2^2\right|<\epsilon\qquad w.p.\qquad 1-e^{-\Omega_{\Delta_{ab},\epsilon}(n)}\tag{27}$$

Proof Since $\mathbb{E}r=0$ and $\|r\|_2^2\le 1$ 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 [23] to conclude that

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

Taking $t := \epsilon n$ then gives (25). For (26) and (27), 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 17 *Under the* $(\mathcal{D}, \gamma, n)$*-stochastic ball model, we have* $D^{(a,b)}1 - D^{(a,a)}1 = 4np + q$*, where*

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

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

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

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

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 $|q_i| \leq (6 + 2\Delta_{ab})n\epsilon$, apply triangle and Cauchy–Schwarz to obtain

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

To finish the argument, apply (25) 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 (27). □

Lemma 18 *Under the* $(\mathcal{D},\gamma,n)$*-stochastic ball model, we have*

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

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

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

where the last step occurs with probability $1-e^{-\Omega_{\Delta_{ab},\epsilon}(n)}$ by a union bound over (26) and (25). □

Lemma 19 *Under the* $(\mathcal{D},\gamma,n)$*-stochastic ball model, we have*

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

Proof Lemma 17 gives

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

Cauchy–Schwarz along with (25) then gives the result. □

Lemma 20 *Under the* $(\mathcal{D},\gamma,n)$*-stochastic ball model, there exists* $C=C(\gamma)$ *such that*

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

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

Proof Fix a and b. Then by Lemma 17, the following holds with probability $1-e^{-\Omega_{m,\Delta_{ab},\epsilon}(n)}$:

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

where the last step is by Cauchy–Schwarz. Taking a union bound with Lemma 18 then gives

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

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

Lemma 21 *Suppose $\epsilon \leq 1$. Then there exists $C = C(\Delta_{ab}, m)$ such that under the $(\mathcal{D}, \gamma, n)$-stochastic ball model, we have*

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

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

Proof First, a quick calculation reveals

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

from which it follows that

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

As such, we have

$$\begin{aligned} \|P_{1^\perp} M^{(a,b)} 1\|_2^2 &= \|P_{1^\perp}(D^{(a,b)} 1 - D^{(a,a)} 1)\|_2^2 \\ &= \|D^{(a,b)} 1 - D^{(a,a)} 1\|_2^2 - \|P_1(D^{(a,b)} 1 - D^{(a,a)} 1)\|_2^2. \end{aligned} \tag{28}$$

To bound the first term, we apply the triangle inequality over Lemma 17:

$$\|D^{(a,b)} 1 - D^{(a,a)} 1\|_2 \leq 4n\|p\|_2 + \|q\|_2 \leq 4n\|p\|_2 + (6 + 2\Delta_{ab})n^{3/2}\epsilon. \tag{29}$$

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

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

With this, we then have

$$\|p\|_2^2 = n\bigg(\frac{1}{n}\sum_{i=1}^n p_i^2 - \mathbb{E}p_1^2 + \mathbb{E}p_1^2\bigg) \leq n\mathbb{E}p_1^2 + n\epsilon \tag{30}$$

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

$$\mathbb{E}p_1^2 = \mathbb{E}\left(\frac{\Delta_{ab}}{2}r_1 + \frac{\Delta_{ab}^2}{4}\right)^2 = \frac{\Delta_{ab}^2}{4}\mathbb{E}r_1^2 + \frac{\Delta_{ab}^4}{16}. \tag{31}$$

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

$$\mathbb{E}r_1^2 \leq \frac{1}{m}. \tag{32}$$

Combining (29), (30), (31) and (32) then gives

$$\|D^{(a,b)}1 - D^{(a,a)}1\|_2 \leq \left(\frac{4n^3\Delta_{ab}^2}{m} + n^3\Delta_{ab}^4 + 16n^3\epsilon\right)^{1/2} + (6+2\Delta_{ab})n^{3/2}\epsilon. \tag{33}$$

To bound the second term of (28), first note that

$$\|P_1(D^{(a,b)}1 - D^{(a,a)}1)\|_2 = \frac{1}{\sqrt{n}}\left|1^\top D^{(a,b)}1 - 1^\top D^{(a,a)}1\right|. \tag{34}$$

Lemma 19 then gives

$$\left|1^\top D^{(a,b)}1 - 1^\top D^{(a,a)}1\right| \geq 1^\top D^{(a,b)}1 - 1^\top D^{(a,a)}1 \geq n^2\Delta_{ab}^2 - (6+4\Delta_{ab})n^2\epsilon \tag{35}$$

with probability $1 - e^{-\Omega_{m,\Delta_{ab},\epsilon}(n)}$. Using (28) to combine (33) with (34) and (35) then gives the result. □

Lemma 22 *There exists $C = C(\gamma)$ such that under the $(\mathcal{D}, \gamma, n)$-stochastic ball model, we have*

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

Proof Recall from (13) that

$$\rho_{(a,b)} = u_{(a,b)}^\top 1 = 1^\top M^{(a,b)}1 - nz = 1^\top M^{(a,b)}1 - n\min_{\substack{a,b\in\{1,\ldots,k\}\\ a\neq b}}\min(M^{(a,b)}1). \tag{36}$$

To bound the first term, we leverage Lemma 19:

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

with probability $1 - e^{-\Omega_{m,\Delta_{ab},\epsilon}(n)}$. To bound the second term in (36), note from Lemma 18 that

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

with probability $1 - e^{-\Omega_{\Delta_{ab},\epsilon}(n)}$. Next, Lemma 17 gives

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

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

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

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

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

Combining these estimates then gives

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

Performing a union bound over a and b then gives

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

Combining these estimates then gives the result. □

Lemma 23 *Under the* $(\mathcal{D}, \gamma, n)$*-stochastic ball model, we have*

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

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

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

and so the triangle inequality gives

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

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 [24]:

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

For the second term, apply (25). The union bound then gives the result. □

Proof of Theorem 9 First, we combine Lemmas 21, 22 and 23: For every $\delta > 0$, there exists an $\epsilon > 0$ such that

$$\begin{aligned}
2\|\Psi\|^2 &+ \sum_{a=1}^{k}\sum_{b=a+1}^{k}\frac{\|P_{1^\perp}M^{(a,b)}1\|_2\|P_{1^\perp}M^{(b,a)}1\|_2}{\rho_{(a,b)}} \\
&\leq 2\bigg(\frac{1+\epsilon}{\sqrt{m}} + \epsilon\bigg)^2 nk + \sum_{a=1}^{k}\sum_{b=a+1}^{k}\frac{4n^3\Delta_{ab}^2/m + Cn^3\epsilon}{n^2(\Delta_{ab}^2 - \Delta(\Delta-2)) - Cn^2\epsilon} \\
&\leq n\bigg(\frac{2k}{m} + \frac{4}{m}\sum_{a=1}^{k}\sum_{b=a+1}^{k}\frac{\Delta_{ab}^2}{\Delta_{ab}^2 - \Delta(\Delta-2)} + \delta\bigg) \qquad (37)
\end{aligned}$$

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

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

Combining this with (37) and considering Lemma 20, 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 $\Delta \geq 2$. □

Appendix 4: Proof of Theorem 14

Put $g = \gamma/\|\gamma\|_2$ and let z have unit 2-norm. Since $\|\Phi_0^\top z\|_2 \geq \|\Phi_0^\top g\|_2$, then considering Lemma 15, it suffices to show that the containment

$$S_1 := \left\{ v \in \mathbb{S}^{m-1} : |\langle g^\top v \rangle| \leq \frac{2}{\Delta} \right\} \subseteq \left\{ v \in \mathbb{S}^{m-1} : \|\Phi_0^\top v\|_2 < \|\Phi_0^\top g\|_2 \right\} =: S_2$$

holds with probability $1 - e^{-\Omega_{m,\Delta}(N)}$. To this end, we will first show that each $v \in S_1$ is also a member of S_2 with high probability, and then we will perform a union bound over an ϵ-net of S_1.

We start by considering $\|\Phi^\top v\|_2$ and $\|\Phi^\top g\|_2$. Decompose x_i as either $\gamma + r_i$ or $-\gamma + r_i$ depending on whether x_i belongs to the ball centered at γ or $-\gamma$. Let w with $\|w\|_2 = 1$ be arbitrary. Then

$$\begin{aligned}(x_i^\top w)^2 &= ((\pm\gamma + r_i)^\top w)^2 = (\pm\gamma^\top w + r_i^\top w)^2 \\ &= (\gamma^\top w)^2 \pm 2(\gamma^\top w)(r_i^\top w) + (r_i^\top w)^2,\end{aligned}$$

and so $\mathbb{E}(x_i^\top w)^2 = (\gamma^\top w)^2 + \mathbb{E}(e_1^\top r)^2$. Linearity of expectation then gives

$$\mathbb{E}\big[(x_i^\top g)^2 - (x_i^\top v)^2\big] = (\gamma^\top g)^2 - (\gamma^\top v)^2 = \|\gamma\|^2(1 - (g^\top v)^2) \geq 1 - \frac{4}{\Delta^2}.$$

Since $|(x_i^\top g)^2 - (x_i^\top v)^2| \leq 2(1 + \Delta/2)^2$ almost surely, we may apply Hoeffding's inequality to get

$$\begin{aligned}\|\Phi^\top g\|_2^2 - \|\Phi^\top v\|_2^2 &= \sum_{i=1}^N \Big((x_i^\top g)^2 - (x_i^\top v)^2 \Big) \\ &\geq N\left(1 - \frac{4}{\Delta^2}\right) - s \quad \text{w.p.} \quad 1 - e^{-\Omega_\Delta(s^2/N)}. \qquad (38)\end{aligned}$$

For a properly chosen t, rearranging gives that $\|\Phi^\top v\|_2 < \|\Phi^\top g\|_2$. Instead, we will use (38) to prove the closely related inequality $\|\Phi_0^\top v\|_2 < \|\Phi_0^\top g\|_2$. Letting μ denote the centroid of the columns of Φ, we know by (25) that $\|\mu\|_2 \leq \delta$ with probability $1 - e^{-\Omega_{m,\delta}(N)}$. In this event, every w with $\|w\|_2 = 1$ satisfies

$$\begin{aligned}\big|\|\Phi_0^\top w\|_2 - \|\Phi^\top w\|_2\big| &= \big|\|(\Phi + \mu 1^\top)^\top w\|_2 - \|\Phi^\top w\|_2\big| \\ &= \big|\|\Phi^\top w + 1\mu^\top w\|_2 - \|\Phi^\top w\|_2\big| \le \|1\mu^\top w\|_2 \le \sqrt{N}\delta.\end{aligned} \tag{39}$$

Furthermore,

$$\|\Phi_0^\top w\|_2 = \|(\Phi - \mu 1^\top)^\top w\|_2 \le \|\Phi w\|_2 + \|1\mu^\top w\|_2 \le \sqrt{N}\bigg(\frac{\Delta}{2} + 1 + \|\mu\|_2\bigg),$$

where the last inequality follows from Cauchy–Schwarz along with the fact that $\|x_i\|_2 \le \Delta/2 + 1$ for every i. Taking a supremum over w then gives

$$\|\Phi_0^\top\|_{2\to 2} \le \sqrt{N}\bigg(\frac{\Delta}{2}+1+\|\mu\|_2\bigg) \le \sqrt{N}\bigg(\frac{\Delta}{2}+1+\delta\bigg) \quad \text{w.p.} \quad 1-e^{-\Omega_{m,\delta}(N)}. \tag{40}$$

In (38), pick $s = (N/2)(1-4/\Delta^2) =: c_1(\Delta)N$. Then taking a union bound with (39) gives

$$\big(\|\Phi_0^\top v\|_2 - \sqrt{N}\delta\big)^2 \le \|\Phi^\top v\|_2^2 \le \|\Phi^\top g\|_2^2 c_1(\Delta)N \le \big(\|\Phi_0^\top g\|_2 + \sqrt{N}\delta\big)^2 - c_1(\Delta)N$$

with probability $1 - e^{-\Omega_{m,\Delta,\delta}(N)}$. Expanding both sides and rearranging then gives

$$\begin{aligned}\|\Phi_0^\top v\|_2^2 &\le \|\Phi_0^\top g\|_2^2 + 2\sqrt{N}\delta\big(\|\Phi_0^\top v\|_2 + \|\Phi_0^\top g\|_2\big) - c_1(\Delta)N \\ &\le \|\Phi_0^\top g\|_2^2 - \underbrace{\bigg(c_1(\Delta) - 4\delta\bigg(\frac{\Delta}{2} + 1 + \delta\bigg)\bigg)}_{c_2(\Delta)} N,\end{aligned}$$

where the last step follows from (40). Thus, picking $\delta = \delta(\Delta)$ sufficiently small ensures $c_2(\Delta) > 0$. Since $c_2(\Delta)N \le \|\Phi_0^\top g\|_2^2 - \|\Phi_0^\top v\|_2^2 = (\|\Phi_0^\top g\|_2 + \|\Phi_0^\top v\|_2)(\|\Phi_0^\top g\|_2 - \|\Phi_0^\top v\|_2)$, we further have

$$\|\Phi_0^\top g\|_2 - \|\Phi_0^\top v\|_2 \ge \frac{c_2(\Delta)N}{\|\Phi_0^\top g\|_2 + \|\Phi_0^\top v\|_2} \ge c_3(\Delta)\sqrt{N},$$

where the last inequality takes $c_3(\Delta) := c_2(\Delta)/(\Delta/2 + 1 + \delta)$, following (40).

At this point, we know that if $v \in S_1$, then $v \in S_2$ with probability $1 - e^{-\Omega_{m,\Delta}(N)}$. It remains to perform a union bound over an ϵ-net of S_1 to conclude that $S_1 \subseteq S_2$ with high probability. To this end, pick $\epsilon < c_3(\Delta)/(\Delta/2 + 1 + \delta)$, consider an ϵ-net $\mathcal{N}_\epsilon$ of S_1, and suppose

$$\|\Phi_0^\top v\|_2 \le \|\Phi_0^\top g\|_2 - c_3(\Delta)\sqrt{N} \qquad \forall v \in \mathcal{N}_\epsilon. \tag{41}$$

Then for every $x \in S_1$, there exists $v \in \mathcal{N}_\epsilon$ such that $\|x - v\|_2 \leq \epsilon$, and so (40) gives

$$\begin{aligned}\|\Phi_0^\top x\|_2 &\leq \|\Phi_0^\top\|\|x - v\|_2 + \|\Phi_0^\top v\|_2\\ &\leq \sqrt{N}\left(\frac{\Delta}{2} + 1 + \delta\right)\epsilon + \|\Phi_0^\top g\|_2 - c_3(\Delta)\sqrt{N} < \|\Phi_0^\top g\|_2,\end{aligned}$$

as desired. To measure the probability of the success event (41), a standard volume comparison argument establishes the existence of an ϵ-net of size $|\mathcal{N}_\epsilon| \leq (1+2/\epsilon)^m$; see Lemma 5.2 in [24]. As such, the union bound gives that (41) occurs with probability $1 - e^{-\Omega_{m,\Delta}(N)}$.

References

1. Abbe, E., Bandeira, A.S., Hall, G.: Exact recovery in the stochastic block model. IEEE Trans. Inf. Theory **62**(1), 471–487 (2016)
2. Abbe, E., Sandon, C.: 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, pp. 670–688, 17–20 October 2015
3. Arthur, D., Vassilvitskii, S.: k-Means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete algorithms (2007)
4. Awasthi, P., Bandeira, A.S., Charikar, M., Krishnaswamy, R., Villar, S., Ward, R.: Relax, no need to round: integrality of clustering formulations. In: Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, pp. 191–200. ACM (2015)
5. Bandeira, A.S.: A note on probably certifiably correct algorithms. C. R. Math. **354**(3), 329–333 (2015)
6. Chen, H., Peng, J.: 0–1 Semidefinite programming for graph-cut clustering: modelling and approximation. In: Data Mining and Mathematical Programming. CRM Proceedings and Lecture Notes of the American Mathematical Society, pp. 15–40 (2008)
7. Dhillon, I.S., Guan, Y., Kulis, B.: Kernel k-means: spectral clustering and normalized cuts. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 551–556. ACM (2004)
8. Dhillon, I.S., Guan, Y., Kulis, B.: Weighted graph cuts without eigenvectors a multilevel approach. IEEE Trans. Pattern Anal. Mach. Intell. **29**(11), 1944–1957 (2007)
9. Elhamifar, E., Sapiro, G., Vidal, R.: Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. In: Advances in Neural Information Processing Systems, pp. 19–27 (2012)
10. Golub, G.H., Van Loan, C.F.: Matrix Computations, vol. 3. JHU Press, Baltimore (2012)
11. Grant, M., Boyd, S., Ye, Y.: Graph implementations for nonsmooth convex programs. In: Blondel, V., Boyd, S., Kimura,H., (eds.) Recent Advances in Learning and Control. Lecture Notes in Control and Information Sciences. Springer, London, pp. 95–110 (2008)
12. Grant, M., Boyd, S.: CVX: matlab software for disciplined convex programming, version 2.1 (2014). http://cvxr.com/cvx
13. Iguchi, T., Mixon, D.G., Peterson, J., Villar, S.: On the tightness of an SDP relaxation of k-means. arXiv preprint arXiv:1505.04778 (2015)
14. Jain, K., Mahdian, M., Saberi, A.: A new greedy approach for facility location problems. In: Proceedings of the 34th Annual ACM Symposium on Theory of Computing (2002)
15. Laurent, B., Massart, P.: Adaptive estimation of a quadratic functional by model selection. Ann. Stat. **28**, 1302–1338 (2000)
16. Lloyd, S.: Least squares quantization in PCM. IEEE Trans. Inf. Theory **28**(2), 129–137 (1982)
17. Mallat, S.G., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Trans. Sig. Process. **41**(12), 3397–3415 (1993)
18. Mixon, D.G.: Cone programming cheat sheet. Short, Fat Matrices (weblog) (2015)
19. Nellore, A., Ward, R.: Recovery guarantees for exemplar-based clustering. Inf. Comput. **245**, 165–180 (2015)
20. Nesterov, Y., Nemirovskii, A.: Interior-Point Polynomial Algorithms in Convex Programming, vol. 13. SIAM, Philadelphia (1994). doi:10.1137/1.9781611970791

21. Ostrovsky, R., Rabani, Y., Schulman, L., Swamy, C.: 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)
22. Peng, J., Wei, Y.: Approximating k-means-type clustering via semidefinite programming. SIAM J. Optim. **18**(1), 186–205 (2007)
23. Tropp, J.A.: User-friendly tail bounds for sums of random matrices. Found. Comput. Math. **12**(4), 389–434 (2012)
24. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. arXiv:1011.3027v7 (2011)
25. Vinayak, R.K., Hassibi, B.: Similarity clustering in the presence of outliers: Exact recovery via convex program. In: IEEE International Symposium on Information Theory, ISIT 2016, Barcelona, Spain, pp. 91–95, 10–15 July 2016
26. Wang, H., Song, M.: Ckmeans.1d.dp: optimal k-means clustering in one dimension by dynamic programming. R J. **3**(2), 29–33 (2011)