Exact Recovery of Mangled Clusters with Same-Cluster Queries (supplementary material)

1 Ancillary results

1.1 VC-dimension of ellipsoids

For any PSD matrix $M$, we denote by $E_M = \{x \in \mathbb{R}^d : d_M(x, \mu) \leq 1\}$ the $\mu$-centered ellipsoid with semiaxes of length $\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2}$, where $\lambda_1, \ldots, \lambda_d \geq 0$ are the eigenvalues of $M$. We recall the following classical VC-dimension bound (see, e.g., [3]).

**Theorem 5.** The VC-dimension of the class $\mathcal{H} = \{E_M : M \in \mathbb{R}^d, M \succeq 0\}$ of (possibly degenerate) ellipsoids in $\mathbb{R}^d$ is $d^2 + 3d + 1$.

1.2 Generalization error bounds

The next result is a simple adaptation of the classical VC bound for the realizable case (see, e.g., [5, Theorem 6.8]).

**Theorem 6.** There exists a universal constant $c > 0$ such that for any family $\mathcal{H}$ of measurable sets $E \subset \mathbb{R}^d$ of VC-dimension $d < \infty$, any probability distribution $\mathcal{D}$ on $\mathbb{R}^d$, and any $\varepsilon, \delta \in (0, 1)$, if $S$ is a sample of $m \geq c d^{d+3d/2}$ points drawn i.i.d. from $\mathcal{D}$, then for any $E^* \in \mathcal{H}$ we have:

$$
\mathcal{D}(E \triangle E^*) \leq \varepsilon \quad \text{and} \quad \mathcal{D}(E' \setminus E^*) \leq \varepsilon
$$

with probability at least $1 - \delta$ with respect to the random draw of $S$, where $E$ is any element of $\mathcal{H}$ such that $E \cap S = E^* \cap S$, and $E'$ is any element of $\mathcal{H}$ such that $E^* \cap S \subseteq E' \cap S$.

The first inequality is the classical PAC bound for the zero-one loss, which uses the fact that the VC dimension of $\{E \triangle E^* : E \in \mathcal{H}\}$ is the same as the VC dimension of $\mathcal{H}$. The second inequality follows immediately from the same proof by noting that, for any $E^* \in \mathcal{H}$ the VC dimension of $\{E \setminus E^* : E \in \mathcal{H}\}$ is not larger than the VC dimension of $\mathcal{H}$ because, for any sample $S$ and for any $F, G \in \mathcal{H}$, $(F \setminus E^*) \cap S \neq (G \setminus E^*) \cap S$ implies $F \cap S \neq G \cap S$.

1.3 Concentration bounds

We recall standard concentration bounds for non-positively correlated binary random variables, see [2]. Let $X_1, \ldots, X_n$ be binary random variables. We say that $X_1, \ldots, X_n$ are non-positively correlated if
correlated if for all \( I \subseteq \{1, \ldots, n\} \) we have:

\[
P(\forall i \in I : X_i = 0) \leq \prod_{i \in I} P(X_i = 0) \quad \text{and} \quad P(\forall i \in I : X_i = 1) \leq \prod_{i \in I} P(X_i = 1)
\]

(1)

**Lemma 4** (Chernoff bounds). Let \( X_1, \ldots, X_n \) be non-positively correlated binary random variables. Let \( a_1, \ldots, a_n \in [0, 1] \) and \( X = \sum_{i=1}^{n} a_i X_i \). Then, for any \( \epsilon > 0 \), we have:

\[
P(X < (1 - \epsilon)E[X]) < e^{-\frac{\epsilon^2 \gamma}{2} E[X]}
\]

(2)

\[
P(X > (1 + \epsilon)E[X]) < e^{-\frac{\epsilon \gamma}{2} E[X]}
\]

(3)

1.4 Yao’s minimax principle

We recall Yao’s minimax principle for Monte Carlo algorithms. Let \( A \) be a finite family of deterministic algorithms and \( I \) a finite family of problem instances. Fix any two distributions \( p \) over \( I \) and \( q \) over \( A \), and any \( \delta \in [0, 1/2] \). Let \( \min_{A \in A} E_{I \sim p} [C_\delta(I, A)] \) be the minimum, over every algorithm \( A \) that fails with probability at most \( \delta \) over the input distribution \( p \), of the expected cost of \( A \) over the input distribution itself. Similarly, let \( \max_{I \in I} E_{A \sim q} [C_\delta(I, A)] \) be the expected cost of the randomized algorithm defined by \( q \) under its worst input from \( I \), assuming it fails with probability at most \( \delta \).

Then (see [4], Proposition 2.6):

\[
\max_{I \in I} E_{q}[C_\delta(I, A)] \geq \frac{1}{2} \min_{A \in A} E_{p}[C_{2\delta}(I, A)]
\]

(4)

2 Supplementary material for Section 5

2.1 Monochromatic Tessellation

We give a formal version of the claim about the monochromatic tessellation of Section 5:

**Theorem 7.** Suppose we are given an ellipsoid \( E \) such that \( \frac{1}{\Phi} E \subseteq \operatorname{conv}(S_C) \subseteq E \) for some stretch factor \( \Phi > 0 \). Then for a suitable choice of \( \beta_i, \rho, b \), the tessellation \( R \) of the positive orthant of \( E \) (Definition 3) satisfies:

(1) \( |R| \leq \max \left\{ 1, O\left(\frac{\phi \Phi d}{\gamma} \ln \frac{\phi \Phi d}{\gamma}\right)^d \right\} \)

(2) \( E \cap \mathbb{R}^d_+ \subseteq \bigcup_{R \in R} R \)

(3) for every \( R \in R \), the set \( R \cap E \) is monochromatic

In order to prove Theorem 7, we define the tessellation and prove properties (1-3) for \( \gamma \leq 1/2 \). For \( \gamma > 1/2 \) the tessellation is defined as for \( \gamma = 1/2 \), and one can check all properties still hold. In the proof we use a constant \( c = \sqrt{5} \) and assume \( \gamma < c^2 - 2c \), which is satisfied since \( c^2 - 2c = 5 - 2\sqrt{5} > 1/2 \).

First of all, we define the intervals \( T_i \). The base \( i \)-th coordinate is:

\[
\beta_i = \frac{\gamma}{c \sqrt{2d} \Phi d}
\]

(5)

Note that, for all \( i \),

\[
\frac{L_i}{\beta_i} = \frac{\Phi c d \sqrt{2d}}{\gamma}
\]

(6)

Define:

\[
\alpha = \frac{\gamma}{c \sqrt{2d} \Phi d}
\]

(7)

and let:

\[
b = \max \left( 0, \log_{1+\alpha} \left( \frac{c \Phi d \sqrt{2d}}{\gamma} \right) \right)
\]

(8)
The parameter $\rho$ of the informal description of Section 5 is exactly $1 + \alpha$). Finally, define the interval set along the $i$-th axis as:

$$T_i = \begin{cases} \{ [0, \beta_i] \} & \text{if } b = 0 \\ \{ [0, \beta_i], (\beta_i, \beta_i(1 + \alpha)], \ldots, (\beta_i(1 + \alpha)^{b-1}, \beta_i(1 + \alpha)^b] \} & \text{if } b \geq 1 \end{cases}$$

(9)

**Proof of (1).** By construction, $|T_i| = b + 1$. Thus, $|R| = \prod_{i \in [d]} |T_i| = (b + 1)^d$. Thus, if $b = 0$ then $|R| = 1$, else by (3) and (5)

\[ b = \frac{\ln(e^{\Phi d \sqrt{2d}})}{\ln(1 + \alpha)} \]

(10)

\[ \leq \frac{2}{\alpha} \ln\left(\frac{e^{\Phi d \sqrt{2d}}}{\gamma}\right) \quad \text{since } \ln(1 + \alpha) \geq \alpha/2 \text{ as } \alpha \leq 1 \]

(11)

\[ = \frac{2\sqrt{2}\Phi d}{\gamma} \ln\frac{e^{\Phi d \sqrt{2d}}}{\gamma} \quad \text{definition of } \alpha \]

(12)

\[ = O\left(\frac{d\Phi}{\gamma} \ln\frac{d\Phi}{\gamma}\right) \quad \text{since } d\Phi \geq 1, \gamma \leq 1/2 \]

(13)

in which case $|R| = O\left(\frac{d\Phi}{\gamma} \ln\frac{d\Phi}{\gamma}\right)^d$. Taking the maximum over the two cases proves the claim.

**Proof of (2).** We show for any $x \in E \cap \mathbb{R}^d$, there exists $R \in \mathcal{R}$ containing $x$. Clearly, if $x \in E \cap \mathbb{R}^d$, then $(x, u_i) \in [0, L_i]$ for all $i \in [d]$. But $T_i$ covers, along the $i$-th direction $u_i$, the interval from $0$ to $\beta_i(1 + \alpha)^b = \beta_i(1 + \alpha)^{\max(0, \log_{1 + \alpha}(L_i/\beta_i))} \geq \beta_i(1 + \alpha)^{\log_{1 + \alpha}(L_i/\beta_i)}$ \geq L_i

(14)

Therefore some $R \in \mathcal{R}$ contains $x$.

**Proof of (3).** Given any hyperrectangle $R \in \mathcal{R}$, we show that the existence of $x, y \in R \cap E$ with $x \in C$ and $y \notin C$ leads to a contradiction. For the sake of the analysis we conventionally set the origin at the center $\mu$ of $E$, i.e. we assume $\mu = 0$.

We define $E_{in} = \frac{1}{\delta^2} E$ and let $M = UU^T$ be its PSD matrix, where $U = [u_1, \ldots, u_d]$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$. Note that $\lambda_i = \frac{1}{\ell_i} = \frac{\Phi^2 d^2}{L_i}$ where $\ell_i = \frac{L_i}{\delta^2}$ is the length of the $i$-th semiaxis of $E_{in}$. For any $R \in \mathcal{R}$, let $R_i$ be the projection of $R$ on $u_i$ (i.e. $R_i$ is one of the intervals of $T_i$ defined in (9)). Let $D = D(R) = \{ i \in [d] : 0 \notin R_i \}$. We let $U_D$ and $U_{-D}$ be the matrices obtained by zeroing out the columns of $U$ corresponding to the indices in $[d] \setminus D$ and $D$, respectively. Observe that if $x, y \in R \cap E$ then:

\[ \langle x - y, u_i \rangle^2 < \alpha^2 \langle x, u_i \rangle^2 \quad \forall i \in D \]

(15)

\[ \langle x - y, u_i \rangle^2 \leq \beta_i^2 \quad \forall i \notin D \]

(16)

Now suppose $C$ has margin at least $\gamma$ for some $\gamma \in (0, c^2 - 2c]$, and suppose $x, y \in R \cap E$ with $x \in C$ and $y \notin C$. Through a set of ancillary lemmata proven below, this leads to the absurd:

\[ \frac{\gamma^2}{c^2} < d_W(y, x)^2 \quad \text{Lemma 5} \]

(17)

\[ \leq d_M(y, x)^2 \quad \text{Lemma 5} \]

(18)

\[ < \alpha^2 d_M(x, \mu)^2 + \frac{\gamma^2}{2c^2} \quad \text{Lemma 7} \]

(19)

\[ \leq \frac{\gamma^2}{2c^2} + \frac{\gamma^2}{2c^2} \quad \text{Lemma 8} \]

(20)

In the rest of the proof we prove the four lemmata.

**Lemma 5.** $\frac{\gamma}{c} < d_W(y, x)$.  

3
Proof. Let \( z \) be the point w.r.t. which the margin of \( C \) holds. By the margin assumption,
\[
d_W(y, z) > \sqrt{1 + \gamma} \quad \text{and} \quad d_W(x, z) \leq 1
\]  
(21)
By the triangle inequality then,
\[
d_W(y, x) \geq d_W(y, z) - d_W(x, z) > \sqrt{1 + \gamma} - 1
\]  
(22)
One can check that for \( \gamma \leq \epsilon^2 - 2\epsilon \) we have \( 1 + \gamma \geq (1 + \frac{\gamma}{\epsilon})^2 \). Therefore
\[
d_W(y, x) > \sqrt{(1 + \gamma/\epsilon)^2} - 1 = \frac{\gamma}{\epsilon}
\]  
(23)
as desired.

Lemma 6. \( d_W(\cdot) \leq d_M(\cdot) \).

Proof. By the assumptions of the theorem, \( E_m \subseteq \text{conv}_M(C) \). Moreover, by the assumptions on \( d_W(\cdot) \), the unit ball of \( d_W(\cdot) \) contains \( \text{conv}(C) \). Thus, the unit ball of \( d_W(\cdot) \) contains the unit ball of \( d_M(\cdot) \). This implies \( W \subseteq M \), thus \( \| \cdot \|_W \leq \| \cdot \|_M \) and \( d_W(\cdot) \leq d_M(\cdot) \).

Lemma 7. \( d_M(y, x)^2 < \alpha^2 d_M(x, \mu)^2 + \frac{\gamma^2}{\epsilon^2} \).

Proof. We decompose \( d_M(y, x)^2 \) along the colspaces of \( U_D \) and \( U_{\sim D} \):
\[
d_M(y, x)^2 = \| M^{1/2}(y - x) \|_2^2 \]
(24)
\[
= \| M^{1/2}(y - x) \|_{U_D U_D^\top}^2 + \| M^{1/2}(y - x) \|_{U_{\sim D} U_{\sim D}^\top}^2
\]  
(25)
Next, we bound the two terms of (25). To this end, we need to show that for all \( D \subseteq [d] \) and \( v \in \mathbb{R}^d \):
\[
\| M^{1/2}v \|_{U_D U_D^\top}^2 = \sum_{i \in D} \lambda_i \langle v, u_i \rangle^2
\]  
(26)
Let indeed \( J_D = \text{diag}(1_D) \) be the selection matrix corresponding to the indices of \( D \). Then \( U_D = U_{J_D} \), and so \( U_{\sim D} U_{\sim D}^\top = J_{\sim D} \). This gives:
\[
\| M^{1/2}v \|_{U_D U_D^\top}^2 = v^\top (U A^{1/2} U^\top) U_D U_D^\top (U A^{1/2} U^\top) v \quad \text{definition of } M \text{ and } \| \cdot \|.
\]  
(27)
\[
= v^\top U A^{1/2} J_D J_D A^{1/2} U^\top v \quad \text{since } U_{D U_D} = J_D
\]  
(28)
\[
= v^\top U J_D A^{1/2} A^{1/2} J_D U^\top v \quad \text{since } A, J_D \text{ are diagonal}
\]  
(29)
\[
= v^\top U_D A U_D^\top v \quad \text{since } U J_D = U_D
\]  
(30)
\[
= \| U_D^\top v \|_A^2 \quad \text{by definition}
\]  
(31)
\[
= \sum_{i \in D} \lambda_i \langle v, u_i \rangle^2
\]  
(32)
Now we can bound the first term of (25):
\[
\| M^{1/2}(y - x) \|_{U_D U_D^\top}^2 = \sum_{i \in D} \lambda_i \langle y - x, u_i \rangle^2
\]  
by (32)
(33)
\[
\leq \alpha^2 \sum_{i \in D} \lambda_i \langle x, u_i \rangle^2 \quad \text{by (15)}
\]  
(34)
\[
= \alpha^2 \| M^{1/2}x \|_{U_D U_D^\top}^2 \quad \text{by (32)}
\]  
(35)
\[
\leq \alpha^2 \| M^{1/2}x \|_{U U^\top}^2
\]  
(36)
\[
= \alpha^2 \| M^{1/2}x \|_2^2
\]  
(37)
\[
= \alpha^2 d_M^2(x, \mu) \quad \text{since } U U^\top = I
\]  
(38)
And for the second term of (25), we have:

\[ \|M_{1/2}(y - x)\|_2^2 \leq \sum_{i \in D} \lambda_i \langle y - x, u_i \rangle^2 \]

by (32) (39)

\[ \leq \sum_{i \in D} \lambda_i \beta_i^2 \]

by (16) (40)

\[ = \sum_{i \in D} \frac{\gamma^2}{2d c^2} \]

by definition of \( \lambda_i \) and \( \beta_i \) (41)

\[ \leq \frac{\gamma^2}{2c^2} \]  

(43)

Summing the bounds on the two terms shows that \( d_M(y, x)^2 < \alpha^2 d_M(x, \mu)^2 + \frac{\gamma^2}{2c^2} \), as claimed.

Lemma 8. \( \alpha^2 d_M(x, \mu)^2 \leq \frac{\gamma^2}{2c^2} \).

Proof. By construction we have \( x \in E \) and \( E = \Phi d \cdot E_{in} \). Therefore \( \frac{1}{\Phi d} x \in E_{in} \), that is:

\[ 1 \geq d_M \left( \frac{1}{\Phi d} x, \mu \right)^2 = \frac{1}{\Phi^2 d^2} d_M(x, \mu)^2 \]  

(44)

where we used the fact that \( d_M(\cdot, \mu)^2 = \| \cdot \|_M^2 \) since \( \mu = 0 \). Rearranging terms, this proves that \( d_M(x, \mu)^2 \leq \Phi^2 d^2 \). Multiplying by \( \alpha^2 \), we obtain:

\[ \alpha^2 d_M(x, \mu)^2 \leq \left( \frac{\gamma}{\sqrt{2c \Phi d}} \right)^2 \Phi^2 d^2 = \frac{\gamma^2}{2c^2} \]  

(45)

as desired.

The proof of the theorem is complete.

2.2 Low-stretch separators and proof of Theorem 3

In this section we show how to compute the separator of Theorem 3. In fact, computing the separator is easy; the nontrivial part is Theorem 3 itself, that is, showing that such a separator always exists.

To compute the separator we first compute the MVEE \( E_J = (M^*, \mu^*) \) of \( S_C \) (see Section 5). We then solve the following semidefinite program:

\[
\max_{\alpha \in \mathbb{R}, \mu \in \mathbb{R}^d, M \in \mathbb{R}^{d \times d}} \alpha \\
\text{s.t. } M \succeq \alpha M^* \\
\langle M, (x - \mu)(x - \mu)^T \rangle \leq 1 \quad \forall x \in S_C \\
\langle M, (y - \mu)(y - \mu)^T \rangle > 1 \quad \forall y \in S_{C^c}
\]

(46)

where, for any two symmetric matrices \( A \) and \( B \), \( \langle A, B \rangle = \text{tr}(AB) \) is the usual Frobenius inner product, implying \( \langle M, (x - \mu)(x - \mu)^T \rangle = d_M(x, \mu)^2 \). In words, the constraint \( M \succeq \alpha M^* \) says that \( E \) must fit into \( E_J \) if we scale \( E_J \) by a factor \( \Phi = \frac{1}{\sqrt{\pi}} \). The other constraints require \( E \) to contain all of \( S_C \) but none of the points of \( S_{C^c} \). The objective function thus minimizes the stretch \( \Phi \) of \( E \).

In the rest of this paragraph we prove Theorem 3.
Proof of Theorem 3 (sketch). To build the intuition, we first give a proof sketch where the involved quantities are simplified. The analysis is performed in the latent space $\mathbb{R}^d$ with inner product $(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top W \mathbf{v}$. Setting conventionally $z = 0$, $C$ then lies in the unit ball $B_0$ and all points of $X \setminus C$ lie outside $\sqrt{1+\gamma} B_0$. For simplicity we assume $\gamma \ll 1$ so that $\sqrt{1+\gamma} \approx 1 + \gamma$, but we can easily extend the result to any $\gamma > 0$. Now fix the subset $S_C \subseteq C$, and let $E_1 = E_1(S_C)$ be the MVEE of $S_C$. Observe the following fact: $B_0$ trivially satisfies (1), but in general violates (2); in contrast, $E_1$ trivially satisfies (2), but in general violates (1). The key idea is thus to “compare” $B_0$ and $E_1$ and take, loosely speaking, the best of the two. To see how this works, suppose for instance $E_1$ has small radius, say less than $\gamma/4$. In this case, $E = E_2$ yields the thesis. Indeed, since the center $\mu^*$ of $E_2$ is in $B_0$, then any point of $E$ is within distance $1 + \gamma/4 \leq \sqrt{1+\gamma}$ of the center of $B_0$, and lies inside $\sqrt{1+\gamma} B_0$. Thus $E_1$ separates $S_C$ from $X \setminus C$, satisfying (1). At the other extreme, suppose $E_1$ is large, say with all its $d$ semiaxes longer than $\gamma/4$. In this case, $E = B_0$ yields the thesis: indeed, by hypothesis $E$ fits entirely inside $\gamma/\gamma E_1$, satisfying (2). Unfortunately, the general case is more complex, since $E_1$ may be large along some axes and small along others. In this case, both $B_0$ and $E_1$ fail to satisfy the properties. This requires us to choose the axes and the center of $E$ more carefully. We show how to do this with the help of Figure 1.

Figure 1: Left: the MVEE $E_1$ of $S_C$ and the affine subspace $U + \mu^*$ (marked simply as $U$) spanned by its largest semiaxes. There is no guarantee that $E_1 \subseteq \sqrt{1+\gamma} B_0$. Right: the separator $E$, centered in the center $\mu$ of $B$, with the largest semiaxis in $U$ and the smallest one in $U_1$. We can guarantee that $S_C \subset E \subset \sqrt{1+\gamma} B_0$.

Let $\{u_1, \ldots, u_d\}$ be the orthonormal basis defined by the semiaxes of $E_1$ and $\ell^*_1, \ldots, \ell^*_d$ be the corresponding semiaxes lengths. We define a threshold $\varepsilon = \gamma^2/d^2$, and partition $\{u_1, \ldots, u_d\}$ as $A_P = \{i : \ell^*_i > \varepsilon\}$ and $A_Q = \{i : \ell^*_i \leq \varepsilon\}$. Thus $A_P$ contains the large semiaxes of $E_1$ and $A_Q$ the small ones. Let $U, U_1$ be the subspaces spanned by $\{u_i : i \in A_P\}$ and $\{u_i : i \in A_Q\}$, respectively. Consider the subset $B = B_0 \cap (\mu^* + U)$. Note that $B$ is a ball in at most $d$ dimensions, since it is the intersection of a $d$-dimensional ball and an affine linear subspace of $\mathbb{R}^d$. Let $\mu$ and $\ell$ be, respectively, the center and radius of $B$. We set the center of $E$ at $\mu$, and the lengths $\ell_i$ of its semiaxes as follows:

$$
\ell_i = \begin{cases} 
\ell^*_i & \text{if } i \in A_P \\
\frac{\ell^*_i}{\sqrt{\varepsilon}} & \text{if } i \in A_Q 
\end{cases}
$$

(47)

Loosely speaking, we are “copying” the semiaxes from either $B_0$ or $E_1$ depending on $\ell^*_i$. In particular, the large semiaxes (in $A_P$) are set so to contain all of $B$ and exceed it by a little, taking care of not intersecting $\sqrt{1+\gamma} B_0$. Instead, the small semiaxes (in $A_Q$) are so small that we can safely set them to $1/\sqrt{\varepsilon}$ times those of $E_1$, so that we add some “slack” to include $S_C$ without risking to intersect $\sqrt{1+\gamma} B_0$. Now we are done, and our low-stretch separator is $(M, \mu)$ where $M = \sum_{i=1}^d \ell^2_i u_i u_i^\top$. This the ellipsoid $E$ that yields Theorem 3. In the next paragraph, we show how we can find efficiently all points in $E$ that belong to $C$.  


2.3 Proof of Theorem 3 (full).

We prove the theorem for $\gamma \leq 1/5$ and use the fact that whenever $C$ has weak margin $\gamma$ then it also has weak margin $\gamma'$ for all $\gamma' > \gamma$. As announced, the analysis is carried out in the latent space $\mathbb{R}^d$ equipped with the inner product $\langle u, v \rangle = u^T W v$. All norms $\|u\|$, distances $d(u, v)$, and (cosine of) angles $\langle u, v \rangle / (\|u\| \|v\|)$ are computed according to this inner product unless otherwise specified. Let $B_0$ be the unit ball centered at the origin, which we conventionally set at $z$, the point in the convex hull of $C$ according to which the margin is computed. Then, by assumption, $C \subseteq B_0$, and $x \notin \sqrt{1+\gamma} B_0$ for all $x \notin C$. For ease of notation, in this proof be denote the MVEE by $E^*$ rather than $E_J$. Let then $(E^*, \mu^*)$ be the MVEE of $S_C$; note that $\mu^* \in \text{conv}(S_C) \subseteq B_0$. We let $u_1, \ldots, u_d$ be the orthonormal eigenvector basis given by the axes of $E^*$ and $\lambda^*_1, \ldots, \lambda^*_d$ the corresponding eigenvalues. Note that if $\min_1 \lambda^*_i \geq 5/\gamma^2$ then $E^*$ has radius $\leq 5/\sqrt{\gamma}$ and thus, since $\mu^* \in B_0$ and $\gamma \leq 1/5$, its distance from $B_0$ is at most $1 + 5/\sqrt{\gamma} = \sqrt{1 + 5\sqrt{\gamma} + \gamma^2/5} < \sqrt{1 + \gamma}$. In this case we can simply set $E = E^*$ and the thesis is proven. Thus, from now on we assume $\min_1 \lambda^*_i \leq 5/\gamma^2$.

![Diagram](https://via.placeholder.com/150)

Figure 2: Left: the separating ball $B_0$ of $C$, the MVEE $E^*$ of $S_C$, and the affine subspace $U + \mu^*$ spanned by its largest semiaxes. Middle: $E$ is our separator centered in the center $\mu$ of the ball $B = U \cap B_0$. Right: a point $x \in S_C$ with its projections onto $U$ and $U_\perp$ with respect to the origin, which we conventionally set at $\mu$ (the center of $E$).

Now let:

$$\varepsilon = \frac{\gamma^3}{32d^2}$$  \hspace{1cm} (48)

and partition (the indices of) the basis $\{u_1, \ldots, u_d\}$ as follows:

$$A_P = \{i \colon \lambda^*_i < 1/\varepsilon^2\}, \quad A_Q = [d] \setminus A_P$$  \hspace{1cm} (49)

Since $\min_1 \lambda^*_i < 5/\gamma^2$ and $5/\gamma^2 \leq 1/\varepsilon^2$, we now define the ellipsoid $E$. Let $U, U_\perp$ be the subspaces spanned by $\{u_i : i \in A_P\}$ and $\{u_i : i \in A_Q\}$ respectively, and let $B = B_0 \cap (\mu^* + U)$. Note that $B$ is a ball, since it is the intersection of a ball and an affine linear subspace. Let $\mu$ and $\ell$ be, respectively, the center and radius of $B$ and define

$$\lambda_i = \begin{cases} (1 - \sqrt{5\gamma/4})\ell^{-2} & i \in A_P \\ \varepsilon \lambda^*_i & i \in A_Q \end{cases}$$  \hspace{1cm} (50)

Then our ellipsoidal separator is $E = \{x \in \mathbb{R}^d : d_M(x, \mu) \leq 1\}$. See Figure 2 for a pictorial representation. We now prove that $E$ satisfies: (1) $S_C \subseteq E$, (2) $E \subseteq \frac{6\sqrt{2d}}{\gamma^2} E^*(S_C)$, (3) $E \subseteq \sqrt{1+\gamma} B_0$. 

7
Proof of (1). Set the center $\mu$ of $E$ as the origin. For all $i \in [d]$ let $U_i = u_i u_i^\top$ and define the following matrices:

$$P_0 = \sum_{i \in A_P} U_i, \quad Q_0 = \sum_{i \in A_Q} U_i$$

$$P = \sum_{i \in A_P} \lambda_i U_i, \quad Q = \sum_{i \in A_Q} \lambda_i U_i$$

$$P_* = \sum_{i \in A_P} \lambda_*^i U_i, \quad Q_* = \sum_{i \in A_Q} \lambda_*^i U_i$$

We want to show that $d_M^2(x, \mu) \leq 1$ for all $x \in S_C$. Note that $d_M^2(x, \mu)$ equals (recall that $\mu = 0$):

$$x^\top P x + x^\top Q x$$

Let us start with the second term of (54). By definition of $Q_*$ and since $\mu_*^\top Q_* = (\mu_* - \mu)^\top Q_* = 0$ because $\mu_* - \mu \in U$,

$$x^\top Q x = \varepsilon x^\top Q_* x = \varepsilon (x - \mu^*)^\top Q_* (x - \mu^*) \leq \varepsilon < \frac{\gamma}{4}$$

where the penultimate inequality follows from $x \in E^*$.

We turn to the first term of (54). If we let $p$, $q$ be the projections of $x - \mu = x$ onto $U, U_\perp$, so that

$$\|p\|^2 = x^\top P_0 x, \quad \|q\|^2 = x^\top Q_0 x$$

then by definition of the $\lambda_i$ we have:

$$x^\top P x = \frac{1 - \sqrt{5\gamma/4}}{\ell^2} \|p\|^2$$

We can thus focus on bounding $\|p\|$. Since $B$ is a ball of radius $\ell$, then $\|p\| \leq \ell + d(p, B)$, where $d(p, B)$ is the distance of $p$ from its projection on $B$—see Figure 3 left.

![Figure 3: Left: a point $x \in S_C \subset B_0$ which lies in $E$ as well. Right: for a fixed $a > 0$, the ratio $b/a$ is maximized when the segment of length $a$ lies on the line passing through the center of $B_0$, in which case $b/a = \frac{\sin \theta}{1 - \cos \theta}$ for some $\theta \in (0, \pi/2)$.]

Now, since $x \in B_0$, the ratio $\frac{d(p, B)}{\|q\|}$ is maximized when $\ell \to 0$ (i.e., $B$ has a vanishing radius), in which case $d(p, B) \leq \sin \theta$ and $\|q\| \geq 1 - \cos \theta$, where $\theta \in (0, \pi/2]$; see Figure 3 right. Then:

$$\frac{\|q\|}{d(p, B)} \geq \frac{1 - \cos \theta}{\sin \theta} = \tan \frac{\theta}{2} \geq \frac{\theta}{2} \geq \frac{\sin \theta}{2} \geq \frac{d(p, B)}{2}$$

where we used the tangent half-angle formula and the Taylor expansion of $\tan \theta$. This yields $d(p, B) \leq \sqrt{2} \|q\|$. Thus:

$$\|p\| \leq \ell + \sqrt{2} \|q\|$$

8
But since $\lambda_i^* \geq 1/\varepsilon^2$ for all $i \in A_Q$:
\[
\|q\|^2 = x^TQ_0x \leq \varepsilon^2 x^TQ_*x = \varepsilon^2(x - \mu^*)^TQ_*(x - \mu^*) \leq \varepsilon^2
\]  
(60)
Therefore:
\[
x^TPx \leq \left(1 - \frac{\sqrt{5\gamma/4}}{\ell^2}\right)\left(\ell + \sqrt{2\varepsilon}\right)^2 \leq \left(1 - \frac{\sqrt{5\gamma/4}}{\ell^2}\right)\left(1 + \sqrt{\varepsilon}/\ell\right)^2
\]  
(61)
Next, we show that $\frac{\sqrt{2\varepsilon}}{\ell} \leq \frac{1}{2}\sqrt{5\gamma/4}$. First,
\[
\sqrt{2\varepsilon} = \sqrt{\frac{\gamma}{32d^2}} = \frac{\gamma\sqrt{\gamma}}{4d}
\]  
(62)
We now temporarily set $\mu^*$ as the origin. We want to show that the projection of $\frac{1}{d}E^*$ on $U$ is contained in $B$. Now, the projection of an ellipsoid on the subspace spanned by a subset of its axes is a subset of the ellipsoid itself, and $U$ is by definition spanned by a subset of the axes of $E^*$. Therefore the projection $P$ of $\frac{1}{d}E^*$ on $U$ satisfies $P \subseteq \frac{1}{d}E^*$. Suppose then by contradiction that $P \nsubseteq B$. Since $\bar{B} = U \cap B_0$, this implies that $\frac{1}{d}E^* \notin B_0$. But by John’s theorem, $\frac{1}{d}E^* \subseteq \text{conv}(S_C)$, and therefore $\text{conv}(S_C) \notin B_0$, which is absurd. Therefore $P \subseteq B$.

Let us get back to the proof, with $\mu$ as the origin. On the one hand, the definitions of $A_P$ and $U$ imply that the largest semiaxis of $E^*$ of length $\ell^* = 1/\sqrt{\min_i \lambda_i^*}$ lies in $U$, thus $P$ has radius at least $\frac{1}{2}\ell^*$. On the other hand $B$ has radius $\ell$, and we have seen that $P \subseteq B$. Therefore, $\ell \geq \frac{1}{2}\ell^*$. Finally, by our assumption on $\min_i \lambda_i^*$, we have $\min_i \lambda_i^* < \gamma^2$ and so $\ell^* > \gamma/\sqrt{d}$. Therefore, $\ell \geq \gamma/\sqrt{d}$, which together with (62) guarantees $\frac{\sqrt{2\varepsilon}}{\ell} \leq \frac{\gamma\sqrt{\gamma}}{4d} = \frac{1}{2}\sqrt{5\gamma/4}$. Thus, continuing (61):
\[
x^TPx \leq (1 - \frac{\sqrt{5\gamma/4}}{\ell})(1 + \frac{1}{2}\sqrt{5\gamma/4})^2
\]  
(63)
Now $(1 - x)(1 + \frac{x}{2}) < 1 - \frac{3}{4}x^2$ for all $x > 0$, thus with $x = \sqrt{5\gamma/4} > \sqrt{\gamma}$ we get:
\[
x^TPx < 1 - \frac{3}{4}\gamma
\]  
(64)
By summing (55) and (64), we get:
\[
x^TPx + x^TQx < 1 - \frac{3}{4}\gamma + \frac{\gamma}{4} < 1
\]  
(65)

**Proof of (2).** Comparing the eigenvalues of $E$ and $E^*$, and using $\ell \leq 1$ and $\gamma \leq 1/\varepsilon$, we obtain:
\[
\frac{\lambda_i}{\lambda_i^*} \geq \begin{cases} \frac{(1 - \sqrt{5\gamma/4}/\ell)^2}{\varepsilon^2} & i \in A_P \\ \varepsilon > \frac{\sqrt{\gamma}}{2} & i \in A_Q \end{cases}
\]  
(66)
Thus the semiaxes lengths of $E$ are at most $\sqrt{2}\varepsilon/\ell$ times those of $E^*$. Now let $E^*_+ \subseteq \text{conv}(S_C)$ be the set obtained by scaling $E^*$ by a factor $2\sqrt{\gamma/\varepsilon} = \frac{64\sqrt{2}\varepsilon/\gamma}$ about its origin $\mu^*$. Note that $\mu^* \in \text{conv}(S_C)$ and, by item (1), $\text{conv}(S_C) \subseteq E$, which implies $\mu^* \in E$. Now, $E^*_+$ contains any set of the form $y + \frac{1}{d}E^*$ if the latter contains $\mu^*$; this includes the set $\frac{\sqrt{2}\varepsilon}{\ell}E^*$ centered in $\mu$, which in turn contains $E$ as we already said.

**Proof of (3).** We prove that $d(x, B_0) < \gamma$ for all $x \in E$. Since $B_0$ is the unit ball, this implies $E \subseteq \sqrt{1 + \gamma} B_0$. Consider then any such $x$. Let again $p, q$ be the projections of $x$ on $U$ and $U_\perp$ respectively. Because $B \subseteq B_0$, $d(x, B_0) \leq d(x, B) \leq d(p, B)^2 + \|q\|^2$. See again Figure 3 left, but with $x$ possibly outside $B_0$. For the first term, note that
\[
d(p, B) \leq \max_{i \in A_P} \sqrt{1/\lambda_i} - \ell
\]  
(67)
By definition of $\lambda_i$, this yields:
\[
d(p, B)^2 \leq \left(\frac{\ell}{\sqrt{1 - \sqrt{5\gamma/4}}} - \ell\right)^2 \leq \left(\frac{1}{\sqrt{1 - \sqrt{5\gamma/4}}} - 1\right)^2
\]  
(because $\ell \leq 1$)
Now we show that the right-hand side is bounded by $\frac{3}{4}\gamma$. Consider $f(x) = \frac{1}{\sqrt{4 \gamma}} - 1$ for $x \in [0, 1/2]$. Now $\frac{d^2 f}{dx^2} = \frac{3}{4} (1 - x)^{-3/2} > 0$, so $f$ is convex. Moreover, $f(1/2) = \sqrt{2} - 1 < 0.83 \cdot 1/2$, and clearly $f(0) = 0 \leq 0.83 \cdot 0$. By convexity then, for all $x \in [0, 1/2]$ we have $f(x) \leq 0.83 x$ which implies $f(x)^2 < 0.75 x^2$. By substituting $x = \sqrt{5 \gamma}/4$, for all $\gamma \leq 1/5$ we obtain:

$$d(p, B)^2 \leq \left( \frac{1}{\sqrt{1 - 5 \gamma/4}} - 1 \right)^2 < \frac{3}{4} \cdot \frac{5}{4} \gamma = \frac{15}{16} \gamma$$ (68)

Let us now turn to $q$. By definition of $Q_0$, of $Q$, and of $\lambda_i$ for $i \in A_Q$, we have:

$$\|q\|^2 = x^T Q_0 x \leq \max_{i \in A_Q} \frac{1}{\lambda_i} x^T Q x = \max_{i \in A_Q} \frac{1}{\lambda_i} x^T Q x$$ (69)

But $x^T Q x \leq 1$ since $x \in E$, and recalling that $\lambda_i^* \geq 1/\varepsilon^2$ for all $i \in A_Q$, we obtain:

$$\|q\|^2 \leq \frac{1}{\varepsilon (1/\varepsilon^2)} = \varepsilon < \frac{\gamma}{16}$$ (70)

Finally, by summing (68) and (70):

$$d(x, B_0)^2 \leq d(p, B)^2 + \|q\|^2 < \gamma$$ (71)

The proof is complete.

3 Supplementary material for Section 6

3.1 Lemma 9

**Lemma 9.** Let $b > 0$ be a sufficiently large constant. Let $S$ be a sample of points drawn independently and uniformly at random from $X$. Let $C = \arg \max_{C_i \in C} |S \cap C_i|$, let $S_C = S \cap C$, and suppose $|S_C| \geq bd^2 \ln k$. If $E$ is any (possibly degenerate) ellipsoid in $\mathbb{R}^d$ such that $S_C = C \cap E$, then with probability at least $1/2$ we have $|C \cap E| \geq |X|/4k$. The same holds if we require that $E \cap (S \setminus S_C) = \emptyset$, i.e., that $E$ separates $S_C$ from $S \setminus S_C$.

**Proof.** Let $n = |X|$ for short, and for any ellipsoid $E$ let $E_X = E \cap X$. We show that, with $C$ defined as above, (i) with probability at least $1 - 1/4$ we have $|C| \geq n/2k$, and (ii) with probability at least $1 - 1/4$, if $|C| \geq n/2k$ then $|E_X \Delta C_i| \leq 1/2|C_i|$ where $\Delta$ denotes symmetric difference. By a union bound, then, with probability at least $1/2$ we have $|E \cap C| \geq |C| - |E_X \Delta C| \geq \frac{1}{2}|C| \geq n/4k$.

(i). Let $B$ be the multiset of samples drawn from $X$, and for every cluster $C_i \in C$ let $N_i$ be the number of samples in $C_i$. Let $s = kbd^2 \ln k$; note that $|S| \leq s$ since there are at most $k$ clusters. Now fix any $C_i$ with $|C_i| < \frac{n}{2k}$. Then $\mathbb{E}[N_i] \leq s |C_i| \leq \frac{k bd^2 \ln k}{2}$, and by standard concentration bounds (Lemma 4 in this supplementary material), we have $\mathbb{P}(N_i \geq bd^2 \ln k) = \exp(-\Omega(b \ln k))$, which for $b$ large enough drops below $1/4k$. Therefore, the probability that $N_i \geq bd^2 \ln k$ when taking $s \leq kbd^2 \ln k$ samples is at most $1/4k$. By a union bound on all $C_i$ with $|C_i| < n/2k$, then, $|C| \geq n/2k$ with probability $1 - 1/4$.

(ii). Consider now any $C_i$ with $|C_i| \geq n/2k$. We invoke the generalization bounds of Theorem 6 in this supplementary material with $\varepsilon = 1/4k$ and $\delta = 1/4k$, on the hypothesis class $\mathcal{H}$ of all (possibly degenerate) ellipsoids in $\mathbb{R}^d$. For $b$ large enough, the generalization error of any ellipsoid $E$ that contains $S_C$ is, with probability at least $1 - 1/4k$, at most $1/4k$, which means $|E_X \Delta C_i| \leq n/4k \leq 1/2|C_i|$, as desired. By a union bound on all clusters, with probability at least $1 - 1/4$ this holds for all $C_i$ with $|C_i| \geq n/2k$. The same argument holds if we require $E$ to separate $S \cap C_i$ from $S \setminus C_i$, see again Theorem 6. By a union bound with point (i) above, we have $E \cap C \leq \frac{1}{2}|C|$ with probability at least $1/2$, as claimed.

3.2 Proof of Lemma 3

Let $X_0 = X$ and $N_0 = n$, and for all $i \geq 1$, let $X_i$ be the set of points not yet labeled at the end of round $i$, let $N_i = |X_i|$, and let $R_i = \mathbb{I}\{N_i \leq N_{i-1}(1 - 1/4k)\}$. Recall that $S_C$ is large enough
so that, by Lemma[7] in this supplementary material, we have \( P(R_i = 1 \mid X_{i-1}) \geq 1/2 \) for all \( i \). For every \( t \geq 1 \) let \( \rho_t = \sum_{i=1}^{t} R_i \). Note that:

\[
N_t \leq N_0 \left( 1 - 1/(4k) \right)^{\rho_t} < ne^{-\frac{\rho_t}{4k}} \tag{72}
\]

If \( \rho_t \geq 4k \ln(1/\varepsilon) \), then \( N_1 < \varepsilon n \) and RECUR\((X, k, \gamma, \varepsilon)\) stops. The number of rounds executed by RECUR\((X, k, \gamma, \varepsilon)\) is thus at most:

\[
r_e = \min\{ t : \rho_t \geq 4k \ln(1/\varepsilon) \}.
\]

Now, for all \( i \geq 1 \) consider the \( \sigma \)-algebra \( F_{i-1} \) generated by \( X_0, \ldots, X_{i-1} \), and define: \( Z_i = R_i - B_i \), where \( B_1, B_2, \ldots \) are Bernoulli random variables where each \( B_i \) has parameter \( 1/(2E[R_i \mid F_{i-1}]) \). Obviously, \( Z_i \leq R_i \), and thus for all \( t \) we deterministically have:

\[
\rho_t = \sum_{i=1}^{t} R_i \geq \sum_{i=1}^{t} Z_i \tag{73}
\]

Now note that:

\[
E[Z_i \mid F_{i-1}] = E[R_i \mid F_{i-1}] \frac{1}{2} E[R_i \mid F_{i-1}] = \frac{1}{2} \tag{74}
\]

Now we can prove the theorem. For the first claim, simply note that \( E[r_e] \leq 8k \ln(1/\varepsilon) \), as this is the expected number of fair coin tosses to get \( 4k \ln(1/\varepsilon) \) heads.

For the second claim, consider any \( t \geq 8k \ln n + 6a \sqrt{k} \ln n \). Letting \( \zeta_t = \sum_{i=1}^{t} Z_i \), the event \( r_0 \geq t \) implies \( \zeta_t < 4k \ln n = \frac{t}{2} - 3a \sqrt{k} \ln n = E[\zeta_t] - \delta \) where \( \delta = 3a \sqrt{k} \ln n \). By Hoeffding’s inequality this event has probability at most \( e^{-2\delta^2/t} \), and one can check that for all \( a \geq 1 \) we have \( 2\delta^2 \geq a \ln n \).

4 Supplementary material for Section[7]

4.1 Proof of Theorem[4]

We state and prove two distinct theorems which immediately imply Theorem[4]:

**Theorem 8.** For all \( 0 < \gamma < 1/\gamma \), all \( d \geq 2 \), and every (possibly randomized) learning algorithm, there exists an instance on \( n \geq 2(\sqrt{1 + \gamma} \frac{d-1}{8\gamma})^d \) points and \( |C| = 3 \) latent clusters such that (1) all clusters have margin \( \gamma \), and (2) to return with probability \( 1/3 \) a clustering \( \hat{C} \) such that \( \Delta(\hat{C}, C) = 0 \) the algorithm must make \( \Omega(n) \) same-cluster queries in expectation.

**Proof.** The idea is the following. We define a single set of points \( X \subset \mathbb{R}^d \) and randomize over the choice of the latent PSD matrix \( W \); the claim of the theorem follows by applying Yao’s minimax principle. Specifically, we let \( X \) be a \( \theta(\sqrt{n}) \)-packing of points on the unit sphere in \( \mathbb{R}^d \). We show that, for \( x \in X \) drawn uniformly at random, setting \( W = (1 + \gamma) \text{diag}(x_1^2, \ldots, x_d^2) \) makes \( x \) an outlier, as its distance \( d_W(x, 0) \) from the origin is \( 1 + \gamma \), while every other point is at distance \( \leq 1 \).

Since there are \( (1/\gamma)^d \) such points \( x \) in our set, the bound follows.

We start by defining the points \( X \) in terms of their entry-wise squared vectors. Consider \( S^+_d = \mathbb{R}^d_+ \cap S_d \) where \( S_d = \{ x \in \mathbb{R}^d : \|x\|_2 \leq 1 \} \) is the unit sphere in \( \mathbb{R}^d \). We want to show that there exists a set of \( \frac{1}{2} (1/\gamma)^d \) points in \( S^+_d \) whose pairwise distance is bigger than \( \varepsilon/2 \), where \( \varepsilon \) is defined before. To see this, recall that the packing number of the unit ball \( B_d = \{ x \in \mathbb{R}^d : \|x\|_2 \leq 1 \} \) is \( M(B_d, \varepsilon) \geq \frac{1}{2} (1/\varepsilon)^d \) — see, e.g., [6]. For \( \varepsilon/2 \) and \( d-1 \), this implies there exists \( Y \subseteq B_{d-1} \) such that \( |Y| \geq (2/\varepsilon)^{d-1} \) and \( \|y - y'\|_2 > \varepsilon/2 \) for all distinct \( y, y' \in Y \). Now, consider the lifting function \( f : B_{d-1} \to \mathbb{R}^d \) defined by \( f(y) = (\sqrt{1 - \|y\|_2^2}, y_1, \ldots, y_{d-1}) \). Define the lifted set \( Z = \{ f(y) : y \in Y \} \). Clearly, every \( z \in Z \) satisfies \( \|z\|_2 = 1 \) and \( z_0 \geq 0 \) so \( z \) lies on the northern hemisphere of the sphere \( S_d \). Moreover, \( \|f(y) - f(y')\|_2 \geq \|y - y'\|_2 \) for any two \( y, y' \in Y \). Hence, we have a set \( Z \) of \( (2/\varepsilon)^{d-1} \) points on the \( d \)-dimensional sphere such that \( \|z - z'\|_2 > \varepsilon/2 \) for all distinct \( z, z' \in Z \). But a hemisphere is the union of \( 2^{d-1} \) orthants, hence some orthant contains at least \( 2^{-(d-1)} (2/\varepsilon)^{d-1} = (1/\varepsilon)^{d-1} \) of the points of \( Z \). Without loss of generality we may assume this is the positive orthant and denote the set as \( Z^+ \).
We now define the input set \( X \subseteq \mathbb{R}^d \) as follows:
\[
X = X^{+} \cup X^{-} = \{ \sqrt{z} : z \in Z^{+} \} \cup \{ -\sqrt{z} : z \in Z^{+} \}
\]
Note that \( n = |X| = 2|Z^{+}| = 2(1/\varepsilon)^d - 1 \). Next, we show how every \( z \in Z^{+} \) defines a clustering instance satisfying the constraints of the theorem. For any \( z^{*} \in Z^{+} \), let \( \mu = (1 + \gamma)z^{*} \) and \( W = \text{diag}(w_1, \ldots, w_d) \), which is PSD as required. Define the following three clusters:
\[
C' = \{ -\sqrt{z} \}, \quad C'' = \{ \sqrt{z} \}, \quad C = X \setminus (C' \cup C'')
\]
where, for \( f : \mathbb{R} \to \mathbb{R}, f(x) = (f(x_1), \ldots, f(x_d)) \). Since \( C' \) and \( C'' \) are singletons, they trivially have weak margin \( \gamma \). We now show that \( C \) has weak margin \( \gamma \) w.r.t. to \( \mu = 0 \); that is, \( d_W(x, \mu)^2 > 1 + \gamma \) for \( x = \pm \sqrt{z} \) and \( d_W(x, \mu)^2 \leq 1 \) otherwise. First, note that \( d_W(x, \mu)^2 = \langle w, x^2 \rangle \).
\[
d_W(x, \mu)^2 = \begin{cases} 
1 + \gamma & \text{if } x \in C', C'' \\
(1 + \gamma)\langle z^{*}, x^2 \rangle & \text{if } x \in C
\end{cases}
\]
(75)
However, by construction of \( Z^{+} \), we have that for all \( x \in C \) and \( z = x^2 \),
\[
\langle z, x^2 \rangle \leq 1 - (\varepsilon/2)^2 = 1 - 2(\langle z, z^{*} \rangle + \langle z^{*}, x^2 \rangle)
\]
which implies \( \langle z^{*}, x^2 \rangle \leq 1 - (\varepsilon/2)^2 / 2 = 1 - \varepsilon^2/8 = 1/(1 + \gamma) \) for \( \varepsilon = \sqrt{8}/(1 + \gamma) \). Therefore (75) gives \( d_W(x, \mu)^2 = (1 + \gamma)\langle z^{*}, x^2 \rangle \leq 1 \). This proves \( C \) has weak margin \( \gamma \) as desired.

The size of \( X \) is:
\[
n \geq 2\left( \frac{1}{8\gamma^{d/(1+\gamma)}} \right) = 2\left( \frac{1 + \gamma}{8\gamma^{d/(1+\gamma)}} \right)^{d-1}
\]
Now the distribution of the instances is defined by taking \( z^{*} \) from the uniform distribution over \( Z^{+} \). Consider any deterministic algorithm running over such a distribution. Note that same-cluster queries always return +1 unless at least one of the two queried points is not in \( C \). As \( C \) contains all points in \( X \) but the symmetric pair \( \sqrt{z^{*}}, -\sqrt{z^{*}} \) for a randomly drawn \( z^{*} \), a constant fraction of the points in \( X \) must be queried before one element of the pair is found with probability bounded away from zero. Thus, any deterministic algorithm that returns a zero-error clustering with probability at least \( \delta \) for any constant \( \delta > 0 \) must perform \( \Omega(n) \) queries. By Yao’s principle for Monte Carlo algorithms then (see Section 1.4 above), any randomized algorithm that errs with probability at most \( 1 - \delta \) for any constant \( \delta > 0 \) must make \( \Omega(n) \) queries as well.

**Theorem 9.** For all \( \gamma > 0 \), all \( d \geq 48(1 + \gamma)^2 \), and every (possibly randomized) learning algorithm, there exists an instance on \( n = \Omega\left( \exp(d/(1 + \gamma)^2) \right) \) points and \( |\mathcal{C}| = 2 \) latent clusters such that (1) all clusters have margin at least \( \gamma \), and (2) to return with probability \( 2/3 \) a clustering \( \hat{C} \) such that \( \Delta(\hat{C}, \mathcal{C}) = 0 \) the algorithm must make \( \Omega(n) \) same-cluster queries in expectation.

**Proof.** We exhibit a distribution of instances that gives a lower bound for every algorithm, and then use Yao’s minimax principle. Let \( p = \frac{1}{2(1+\gamma)} \). Consider a set of vectors \( x_1, \ldots, x_n \) where every entry of each vector \( x_{j,i} \) is i.i.d. and it is equal to 1 with probability \( p \). Define \( X = \{ x_1, \ldots, x_n \} \); note that in general \( |X| \leq n \) since the points might not be all distinct. Let \( x^{*} = x_n, \mathcal{C} = \{ x_1, \ldots, x_{n-1} \}, \mathcal{C} = \{ x^{*} \} \). The latent clustering is \( \mathcal{C} = \{ \mathcal{C}, \mathcal{C}' \} \), and the matrix and center of \( \mathcal{C} \) are respectively \( W = \text{diag}(x^{*}) \) and \( c = 0 \). The algorithms receive in input a random permutation of \( X \); clearly, if it makes \( o(|X|) \) queries, then it has vanishing probability to find \( x^{*} \), which is necessary to return the latent clustering \( \mathcal{C} \).

Now we claim that, if \( d \geq 48(1 + \gamma)^2 \), then we can set \( n = \Omega\left( \exp\left(\frac{d}{48(1+\gamma)^2}\right) \right) \) and with constant probability we will have (i) \( |X| = \Omega(n) \), and (ii) \( \mathcal{C}, \mathcal{C}' \) have margin \( \gamma \). This is sufficient, since the theorem then follows by applying Yao’s minimax principle.

Let us first bound the probability that \( |X| < n \). Note that for any two points \( x_i, x_i' \) with \( i \neq i' \) we have \( P(x_i = x_i') = (1 - p)^2 + p^2 d < (1 - 1/2^{(1+\gamma)})^d < e^{-\frac{d}{2^{(1+\gamma)}}} \). Therefore, by a simple union bound over all pairs, \( P(|X| < n) < n^2 e^{-\frac{d}{2^{(1+\gamma)}}} \).
Next, we want show that, loosely speaking, \(d_W(x, c)^2 \simeq dp\) for \(x \in C\) whereas \(d_W(x, c)^2 \simeq dp^2\) for \(x \in C\); this will give the margin.

Now, for any \(x\),
\[
d_W(x, c)^2 = \sum_{i=1}^{d} x_i^*(x_i - 0)^2 = \begin{cases} \sum_{i=1}^{d} x_i^* x_i \sim B(d, p^2) & x \in C \\ \sum_{i=1}^{d} x^* \sim B(d, p) & x \in C' \end{cases}
\]

Where in the last equality we use the fact that the entries are unary, and where with the notation we have an accurate enough estimate of \(\mu_1\) where the penultime inequality holds since Theorem 8 above.

The rest of the proof and the application of Yao’s principle is essentially identical to the proof of Theorem 8 above.

5 Comparison with SCQ-\(k\)-means

In this section we compare our algorithm to SCQ-\(k\)-means of [1]. We show that, in our setting, SCQ-\(k\)-means fails even on very simple instances, although it can still work under (restrictive) assumptions on \(\gamma\), \(W\), and the centers.

SCQ-\(k\)-means works as follows. First, the center of mass \(\mu_c\) of some cluster \(C\) is estimated using \(O(\text{poly}(k, 1/\gamma))\) SCQ queries; second, all points in \(X\) are sorted by their distance from \(\mu_c\) and the radius of \(C\) is found via binary search. The binary search is done using same-cluster queries between the sorted points and any point already known to be in \(C\). The margin condition ensures that, if we have an accurate enough estimate of \(\mu_c\), then the binary search will be successful (there are no inversions of the sorted points w.r.t. their cluster). This approach thus yields a \(O(\ln n)\) SCQ queries bound (the number of queries to estimate \(\mu_c\) is independent of \(n\)).

It is easy to see that this algorithm relies crucially on (1) each cluster \(C\) must be spherical, and (2) the center of the sphere must coincide with the centroid \(\mu_c\). In formal terms, the setting of [1] is a special cases of ours where for all \(C\) we have \(W_C = I_d\) and \(c = E_{x \in C}[x]\). If any of these two assumptions does not hold, then it is easy to construct instances where [1] fails to recover the clusters and, in fact, achieves error very close to a completely random labeling. Formally:
Lemma 10. For any fixed $d \geq 2$, any $p \in (0, 1)$, and any sufficiently small $\gamma > 0$, there are arbitrarily large instances on $n$ points and $k = 2$ clusters on which SCQ-$k$-means incurs error $\Delta(\hat{C}, C) \geq \frac{1-p}{2}$ with probability at least $1 - p$.

Sketch of the proof. We describe the generic instance on $n$ points for $d = 2$. The latent clustering $C$ is formed by two clusters $C_1, C_2$ of size respectively $n_1 = n^{1+\frac{1}{2}}$ and $n_2 = n^{1-\frac{1}{2}}$. In $C_1$, half of the points are in $(1, 0)$ and half in $(-1, 0)$. In $C_2$, all points are in $(0, \sqrt{\frac{1}{2}})$. (One can in fact perturb the instance so that all points are distinct without impairing the proof). For both clusters, the center coincide with their center of mass $\mu_1 = (0, 0)$ and $\mu_2 = (0, \sqrt{\frac{1}{2}})$. For both clusters, the latent metric is given by the PSD matrix $W = (\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix})$. It is easy to see that $d_W(x, \mu_1)^2 = 1/4$ if $x \in C_1$ and $d_W(x, \mu_1)^2 = (1+\gamma)/4$ if $x \in C_2$, and so $C_1$ has margin exactly $\gamma$. On the other hand $C_2$ has margin $\gamma$ since $d_W(x, \mu_2)^2 = 0$ if $x \in C_2$ and $d_W(x, \mu_2)^2 > 0$ otherwise.

![Figure 4: A bad instance for SCQ-$k$-means. With good probability, the algorithm classifies all points in a single cluster, incurring error $\approx 1/2$, the same as a random labeling.](image)

Next, we show that the approach [1] still works if one relaxes the assumption $W = I$, at the price of strengthening the margin $\gamma$. Let $\lambda_{\max}$ and $\lambda_{\min} > 0$ be, respectively, the largest and smallest eigenvalues of $W$. The condition number $\kappa_W$ of $W$ is the ratio $\lambda_{\max}/\lambda_{\min}$. If $\kappa_W$ is not too large, then $W$ does not significantly alter the Euclidean metric, and the ordering of the points is preserved. Formally:

Lemma 11. Let $\kappa_W$ be the condition number of $W$. If every cluster $C$ has margin at least $\kappa_W - 1$ with respect to its center of mass $\mu_C$, and if we know $\mu_C$, then we can recover $C$ with $O(\ln n)$ SCQ queries.

Proof. Fix any cluster $C$ and let $\mu = \mu_C$. For any $z \in \mathbb{R}^d$ we have $\lambda_{\min} \|z\|_W^2 \leq \|z\|_W^2 \leq \lambda_{\max} \|z\|_W^2$ where $\lambda_{\min}$ and $\lambda_{\max}$ are, respectively, the smallest and largest eigenvalue of $W$. Sort all other points $x$ by their Euclidean distance $\|x - \mu\|_2$ from $\mu$. Then, for any $x \in C$ and any $y \not\in C$ we have:

$$\frac{\|y - \mu\|_W^2}{\|x - \mu\|_W^2} \geq \frac{\lambda_{\min}}{\lambda_{\max}} \frac{\|y - \mu\|_W^2}{\|x - \mu\|_W^2} = \frac{1}{\kappa_W} \frac{d(y, \mu)^2}{d(x, \mu)^2} > \frac{1 + \gamma}{\kappa_W} \tag{83}$$

Hence, if $\gamma \geq \kappa_W - 1$, there is $r > 0$ such that $\|x - \mu\|_2 \leq r$ for all $x \in C$ and $\|y - \mu\|_2 \geq r$ all $y \not\in C$. We can thus recover $C$ via binary search as in [1].

As a final remark, we observe that the above approach is rather brittle, since $\kappa_W$ is unknown (because $W$ is), and if the condition $\kappa_W \leq 1 + \gamma$ fails, then once again the binary search can return a clustering far from the correct one.
6 Comparison with metric learning

In this section we show that metric learning, a common approach to latent cluster recovery and related problems, does not solve our problem even when combined with same-cluster and comparison queries. Intuitively, we want to learn an approximate distance $d$ that preserves the ordering of the distances between the points. That is, for all $x, y, z \in X$, $d(x, y) \leq d(x, z)$ implies $\tilde{d}(x, y) \leq \tilde{d}(x, z)$. If this holds then $d$ and $\tilde{d}$ are equivalent from the point of view of binary search. To simplify the task, we may equip the algorithm with an additional comparison query CMP, which takes in input two pairs of points $x, x'$ and $y, y'$ from $X$ and tells precisely whether $d(x, x') \leq d(y, y')$ or not. It turns out that, even with SCQ+CMP queries, learning such a $d$ requires to query essentially all the input points.

**Theorem 10.** For any $d \geq 3$, learning any $\tilde{d}$ such that, for all $x, y, z \in X$, if $d(x, y) \leq d(x, z)$ then $\tilde{d}(x, y) \leq \tilde{d}(x, z)$, requires $\Omega(n)$ SCQ+CMP queries in the worst case, even with an arbitrarily large margin $\gamma$.

**Proof.** We reduce the problem of learning the order of pairwise distances induced by $W$, which we call ORD, to the problem of learning a separator hyperplane, which we call SEP and whose query complexity is linear in $n$.

Problem SEP is as follows. The inputs are a set $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ (the observations) and a set $\mathcal{H} = \{h_1, \ldots, h_k\} \subseteq \mathbb{R}^d$ (the hypotheses). We require that $h_i \in \mathbb{R}^d$. We have oracle access to $\sigma : X \rightarrow \{+1, -1\}$ such that $\sigma(h) = \text{sgn}(\langle h, \cdot \rangle)$ for some $h \in \mathcal{H}$. The output is the $h \in \mathcal{H}$ that agrees with $\sigma$. We assume $\mathcal{H}, X$ support a margin: $\exists \varepsilon > 0$, possibly dependent on the instance, such that $\text{sgn}(\langle h, x \rangle) = \text{sgn}(\langle h, x' \rangle)$ for all $x'$ with $\|x - x'\| \leq \varepsilon$. (Note that this is not the cluster margin $\gamma$).

Let $Q_{\text{ORD}}(n)$ and $Q_{\text{SEP}}(n)$ be the query complexities of ORD and SEP on $n$ points. We show:

**Lemma 12.** $Q_{\text{ORD}}(3n) \leq Q_{\text{SEP}}(n)$.

**Proof.** Let $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ be the input points for SEP and let $h \in \mathbb{R}^d$ be the target hypothesis. By scaling the dataset we can assume $\|x_i\| \leq \varepsilon$ for any desired $\varepsilon$ (even dependent on $n$). We define an instance of ORD on $n' = 3n$ points as follows. First, $W = \text{diag}(h)$. Second, the input set is $X' = S_1 \cup \ldots \cup S_n$ where for $i = 1, \ldots, n$ we define $S_i = \{a_i, b_i, c_i\}$ with:

\begin{align*}
a_i &= 6^i \cdot 1 \\
b_i &= 2 \cdot a_i \\
c_i &= 3 \cdot a_i + x_i
\end{align*}

We first show that a solution to ORD gives a solution of SEP. Suppose indeed that for all pairs of points $\{q, p\}, \{x, y\}$ we know whether $d_W(q, p) \leq d_W(x, y)$. This is equivalent to knowing the output of $\text{CMP}(\{q, p\}, \{x, y\})$, which is

\[\text{CMP}(\{q, p\}, \{x, y\}) = \text{sgn} (\langle h, (q - p)^2 - (x - y)^2 \rangle)\]

Consider then the point $q = c_i, p = x = b_i, y = a_i$ for each $i$. Then:

\[\text{CMP}(\{q, p\}, \{x, y\}) = \text{sgn} (\langle h, (a_i - b_i)^2 - (b_i - c_i)^2 \rangle\]

\[
= \text{sgn} (\langle h, (a_i)^2 - (-a_i - x_i)^2 \rangle) \\
= \text{sgn} (\langle h, 2 \cdot 6^i x_i - x_i^2 \rangle) \\
= \text{sgn} (\langle h, x_i \left(1 - \frac{x_i}{2 \cdot 6^i} \right) \rangle)
\]

By the margin hypothesis, for $\varepsilon$ small enough this equals $\text{sgn}(\langle h, x_i \rangle)$, i.e., the label of $x_i$ in SEP.

We now show that all the other queries reveal no information about the solution of SEP. Suppose then the points are not in the form $q = c_i, p = x = b_i, y = a_i$. Without loss of generality, we can assume that $q > p$ and $q \geq x > y$. It is then easy to see that, for $\varepsilon$ small enough, $(q - p)^2 - (x - y)^2 > 0$ or $(q - p)^2 - (x - y)^2 < 0$. This holds independently of the $x_i$ and of $W$ and therefore gives no information about the solution of SEP.

It follows that, if we can solve ORD in $f(3n)$ CMP queries, then we can solve SEP in $f(n)$ queries. Finally, note that adding SCQ queries does not reduce the query complexity (e.g., let $X$ lie in a single cluster). For the same reason, we can even assume an arbitrarily large cluster margin $\gamma$. \qed
It remains to show that SEP requires $\Omega(n)$ CMP queries in the worst case. This is well known, but we need to ensure that $\mathcal{H} \subset \mathbb{R}^d_+$ and that any $h \in \mathcal{H}$ supports a margin as described above.

Consider the following set $X = \{ x_1, \ldots, x_n \} \subseteq \mathbb{R}^3$:

$$x_i = (1 - \delta, -\cos(\theta_i), -\sin(\theta_i)) \quad (92)$$

where $\theta_i = i \frac{\pi}{2n}$ and $\delta$ is sufficiently small. Let $\mathcal{H} = \{ h_1, \ldots, h_n \}$, where

$$h_j = (1, \cos(\theta_j), \sin(\theta_j)) \quad (93)$$

Note that $\mathcal{H} \subset \mathbb{R}^d_+$ as required. Clearly:

$$\langle h_j, x_i \rangle = \begin{cases} -\delta & \text{if } j = i \\ 1 - (\delta + \cos(\theta_i - \theta_j)) & \text{if } j \neq i \end{cases} \quad (94)$$

By choosing $\delta = \frac{1 - \cos(\pi/2n)}{2}$ we have $\text{sgn} \langle h, x_i \rangle = -1$ if and only if $i = j$. Clearly, any algorithm needs to probe $\Omega(n)$ labels to learn $h$ with constant probability for some $h \in \mathcal{H}$. Finally, note that any $h$ supports a margin, as required.

\[ \square \]

References

[1] Hassan Ashtiani, Shrinu Kushagra, and Shai Ben-David. Clustering with same-cluster queries. In Advances in Neural Information Processing Systems 29, pages 3216–3224, 2016.

[2] Devdatt Dubhashi and Alessandro Panconesi. Concentration of Measure for the Analysis of Randomized Algorithms. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[3] Hervé Fournier and Olivier Teytaud. Lower bounds for comparison based evolution strategies using vc-dimension and sign patterns. Algorithmica, 59(3):387–408, March 2011.

[4] Rajeev Motwani and Prabhakar Raghavan. Randomized Algorithms. Cambridge University Press, USA, 1995.

[5] Shai Shalev-Shwartz and Shai Ben-David. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, USA, 2014.

[6] Roman Vershynin. High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.