PROBABILISTIC ANALYSIS OF SOME EUCLIDEAN CLUSTERING PROBLEMS

A.M. FRIEZE

Dept. of Computer Science and Statistics, Queen Mary College, Mile End Road, London, E1 4NS, England

Received 28 May 1979
Revised 2 July 1980

We are given \( n \) points distributed randomly in a compact region \( D \) of \( R^m \). We consider various optimisation problems associated with partitioning this set of points into \( k \) subsets. For each problem we demonstrate lower bounds which are satisfied with high probability. For the case where \( D \) is a hypercube we use a partitioning technique to give deterministic upper bounds and to construct algorithms which with high probability can be made arbitrarily accurate in polynomial time for a given required accuracy.

1. Introduction

We are given \( n \) points \( X = \{ x^{(1)}, \ldots, x^{(n)} \} \) belonging to a given compact region \( D \subseteq R^m \). We study in this paper various optimisation problems associated with such a set:

**Problem 1.** Find \( Y = \{ y^{(1)}, \ldots, y^{(k)} \} \subseteq D \) such that

\[
    z_1(X, Y) = \max(\min(\|x^{(i)} - y^{(j)}\| : j = 1, \ldots, k) : i = 1, \ldots, n)
\]

is minimised.

**Problem 2.** Find \( Y = \{ y^{(1)}, \ldots, y^{(k)} \} \subseteq X \) such that \( z_1(X, Y) \) is minimised. It will be convenient to refer to the objective function as \( z_2(X, Y) \) in this case.

**Problem 3.** Partition \( X \) into \( k \) subsets \( X_1, \ldots, X_k \) so that

\[
    z_3(X_1, \ldots, X_k) = \max(\max(\|x - y\| : x, y \in X_j) : j = 1, \ldots, k)
\]

is minimised.

**Problem 4.** Partition \( X \) into \( k \) subsets \( X_1, \ldots, X_k \) so that

\[
    z_4(X_1, \ldots, X_k) = \max\left( \sum_{x, y \in X_j} \|x - y\| : j = 1, \ldots, k \right)
\]

is minimised.
The norms considered will be
\[ \|x\|_c = \left( \sum_{j=1}^{m} x_j^2 \right)^{1/2}, \]
\[ \|x\|_\infty = \max(|x_i| : j = 1, \ldots, m). \]
Non-euclidean versions of the above problems are known to be NP-hard as are the corresponding problems of finding \( \varepsilon \)-optimal solutions for arbitrary \( \varepsilon > 0 \).

(For \( m = 1 \) problems 1, 2, 3 are solvable in polynomial time using dynamic programming, the status of problem 4 when \( m = 1 \) is not known.)

It is likely therefore that problems 1–4 are also NP-hard as is the case for Euclidean versions of other NP-hard problems [2, 3]. This paper conducts a probabilistic analysis of these problems. The \( n \) points are assumed to be randomly and uniformly distributed over the region \( D \) which is assumed to have hyper-volume \( V \).

Results can be obtained for other norms by using the fact that for any two norms \( \| \|_a, \| \|_b \) there exists a constant \( p \) such that for \( x \in R^n \|x\|_a \leq p \|x\|_b \). For example if \( m = 2 \) and \( k \), \( n \) grow so that \( k/n \to 0 \) as \( n \to \infty \) we show that in problem 1 using \( \| \|_c \) that
\[ z_1^* = \min z_1(X, Y) \geq (V/k\pi)^{1/2} \]
with probability tending to 1. Now as \( \|x\|_\infty \geq \|x\|_c/\sqrt{2} \) this implies that using \( \| \|_\infty \)
\[ z_1^* \geq (V/2k\pi)^{1/2} \]
with probability tending to 1. We can however prove in this case that \( z_1^* \geq \frac{1}{4}(V/k)^{1/2} \) with probability tending to 1. We have thus analysed these norms separately.

We follow the approach used in Fisher and Hochbaum [1]. For an instance of problem \( t \) we denote the value of an optimal solution by \( z_t^*(n, k) \). For each problem we derive lower bounds for \( z_t^* \) which are valid with probability tending to 1 assuming that \( k/n \to d < 1 \) in problems, 1, 2, 3 and \( d \leq \frac{1}{2} \) for problem 4.

Then restricting our attention to the case where \( D \) is a hypercube we derive simple upper bounds for \( z_t^* \). We then use a grid technique as in Fisher and Hochbaum [1] such that given \( \varepsilon > 0 \) we derive a solution of value \( \hat{z}_t \) where \( \hat{z}_t - z_t^* \leq \varepsilon z_t^* \) with probability tending to 1. The time complexity of these algorithms are \( O(n^{p(\varepsilon)}) \) where \( p(\varepsilon) \) naturally depends on \( \varepsilon \). Fisher and Hochbaum analysed the \( k \)-median problem: find \( Y = \{y^{(1)}, \ldots, y^{(k)}\} \subseteq X \) such that
\[ \sum_{i=1}^{n} \min(\|x^{(i)} - y^{(j)}\| : i = 1, \ldots, k) \]
is minimised.

They only considered \( m = 2 \) and \( \| \|_c \) but their analysis would extend easily to general \( m \).
The results obtained here can be usefully compared with those of [1], most importantly for problem 2 with \( m = 2 \) and \( \| \|_e \) we show that for a fixed region the optimal value (usually) grows like \( 1/\sqrt{k} \) whereas for the \( k \)-median problem the optimal value grows like \( n/\sqrt{k} \). The factor \( n \) is what one would expect on comparing objective functions.

2. Analysis of problem 1

We first compute a probabilistic lower bound to problem 1 using \( \| \|_e \). We shall use Stirling’s inequalities

\[
(n/e)^n (2n\pi)^{1/2} \leq n! \leq (12n/12n-1)(n/e)^n (2n\pi)^{1/2}
\]

several times to replace factorials and so we have stated them here for convenience.

**Notation.** For \( a \in R \), \( a \geq 0 \) and \( c \in R^m \) the hypersphere is

\[
HS(c, a) = \{ x \in R^m : \| x - c \|_e \leq a \}.
\]

It’s hypervolume is denoted by \( c_m a^m \) where the \( c_m \) satisfy

\[
c_1 = 2 \text{ and } c_{m+1} = \left( 2 \int_0^{\pi/2} \cos^{m+1} \theta \, d\theta \right) c_m \text{ for } m \geq 1.
\]

Note that

\[
\int_0^{\pi/2} \cos^{2n} \theta \, d\theta = \frac{2n-1}{2n} \cdot \frac{2n-3}{2n-2} \cdots \frac{1}{2},
\]

\[
\int_0^{\pi/2} \cos^{2n+1} \theta \, d\theta = \frac{2n}{2n+1} \cdot \frac{2n-2}{2n-1} \cdots \frac{2}{3}.
\]

Let \( X \subseteq R^m \) be finite. Let \( r = r(X) \) be the radius of the smallest hypersphere containing \( X \) and let \( c = c(X) \) be the centre of this hypersphere.

**Lemma 2.1.** Let \( X, r, c \) be as above and suppose \( r \leq a \). If \( z \in R^m \) is such that \( r(X \cup \{z\}) \leq a \), then \( \| z - c \|_e \leq a + (a^2 - r^2)^{1/2} \).

**Proof.** Let \( Y = \{ x \in X : \| x - c \|_e = r \} \neq \emptyset \). Now let \( C = \text{convex hull of } Y \). We show by contradiction that \( c \in C \). If \( c \notin C \) let \( b \) be the nearest point of \( C \) to \( c \). Let \( c_\lambda = (1 - \lambda) c + \lambda b \) for \( 0 < \lambda < 1 \). Now for \( \lambda > 0 \) and \( y \in Y \| y - c_\lambda \|_e < \| y - c \|_e \) and so if \( \lambda \) is “small enough” \( c \) can be the centre of a hypersphere of radius \( r' < r \) containing \( X \). Thus \( c \in C \) and so \( c = \sum_{i=0}^d \lambda_i y_i \) where \( \lambda_i > 0 \) for \( i = 1, \ldots, d \) and \( \sum_{i=1}^d \lambda_i = 1 \). Now let \( z \) be as in the statement of the lemma and let \( c_z = c(X \cup \{z\}) \).
Since $c \in C$ there exists $y$, such that $(c \cdot y) \cdot (y \cdot c) \leq 0$. Then

$$a^2 \geq \|c - y\|^2_c$$

$$= \|c\|^2_c + \|y\|^2_c - 2(c \cdot y) \cdot (y \cdot c)$$

$$\geq \|c\|^2_c + r^2.$$ 

Thus $\|c - y\| \leq (a^2 - r^2)^{1/2}$ and hence

$$\|z - c\| \leq \|z - c\| + \|c - y\| \leq a + (a^2 - r^2)^{1/2}. \quad \Box$$

**Lemma 2.2.** Let $E(n, c, a)$ be the event that $n$ points $x^{(1)}, \ldots, x^{(n)}$ chosen at random in $D$ lie in $HS(c, a)$ and let $F(n, a) = \bigcup_{c \in D} (n, c, a)$. Then for $n \geq 2$ there exists $b = b(m) > 1$ such that

$$P(n, a) = \text{Prob}(F(n, a)) \leq b^{\sqrt{n}v^{n-1}} \quad (2.1)$$

where $v = c_m a^m / V$.

**Proof.** Let $p(n, z)$ be the density function of the random variable $z = \text{radius of the smallest hypersphere containing } n \text{ random points in } D$.

It follows that $P(n, a) = \int_0^a p(n, z) \, dz$ and it follows from Lemma 2.1 that

$$P(n + 1, a) \leq \left( \frac{c_m}{V} \right) \int_0^a p(n, z) (a + (a^2 - z^2)^{1/2})^m \, dz.$$ 

Integration by parts gives

$$P(n + 1, a) \leq \left( \frac{c_m a^m}{V} \right) P(n, a)$$

$$+ \left( \frac{mc_m}{V} \right) \int_0^a P(n, z) (a + (a^2 - z^2)^{1/2})^{m-1} z(a^2 - z^2)^{-1/2} \, dz. \quad (2.2)$$

Now $P(2, a) \leq 2^m c_m a^m / V$ and if for some $n \geq 2$ and constant $\alpha$ $P(n, z) \leq \alpha(c_m z^m / V)^{n-1}$ for all $z \geq 0$ then substitution in (2.2) gives

$$P(n + 1, a) \leq \alpha (1 + u_{n-1}) (c_m a^m / V)^n$$

where

$$u_{n-1} = m \int_0^{\pi/2} (1 + \cos \theta)^{n-1} (\sin \theta)^{m(n-1)+1} \, d\theta.$$ 

We deduce therefore that for $n \geq 2$

$$P(n, a) \leq 2^m \prod_{i=1}^{n-2} (1 + u_i) (c_m a^m / V)^{n-1}. \quad (2.3)$$

The RHS of (2.3) is bounded by $b^{\sqrt{n}v^{n-1}}$ for some $b$ dependent on $m$. This can be
shown as follows:

\[ u_n \leq m 2^{n-1} \left[ \frac{\pi}{2} \right] (\sin \theta)^{m+1} d\theta = \alpha \frac{mn}{mn+1} \cdot \frac{mn-2}{mn-1} \ldots \]

where \( \alpha \) is dependent on \( m \). If

\[
\beta_M = \frac{M}{M+1} \cdot \frac{M-2}{M-1} \cdot \frac{2}{3} \quad \text{for } M \text{ even,}
\]

\[
= \frac{M}{M+1} \cdot \frac{M-2}{M+1} \cdot \frac{1}{2} \quad \text{for } M \text{ odd,}
\]

we show that \( \beta_M < 2/\sqrt{M} \). For \( M \) even

\[
\beta_M < 1 \cdot \frac{M-1}{M} \cdot \frac{M-3}{M-2} \cdot \frac{3}{4}
\]

Thus \( \beta_M^2 < 2/(M+1) < 2/M \). For \( M \) odd \( \beta_M < 1/M \) by a similar argument. Thus

\[ u_n < 2\alpha/\sqrt{mn} = \beta/\sqrt{n} \quad \text{for } \beta = 2\alpha/\sqrt{m}. \]

Thus for \( n \geq 2 \)

\[ \prod_{i=1}^{n-2} (1 + u_i) \leq \prod_{i=1}^{n-2} (1 + \beta/\sqrt{i}) \leq e^{2\beta \sqrt{n-2}} \]

as may be shown by induction on \( n \).

Thus the R.H.S. of (2.3) \( < 2^n e^{2\beta \sqrt{n-2}} v^{n-1} \) which can be simplified to \( b^{\sqrt{n}} v^{n-1} \) for large enough \( b \).

**Lemma 2.3.** Let \( n \) points \( X = \{x^{(1)}, \ldots, x^{(n)}\} \) be chosen at random in \( D \). For \( a \geq 0 \) and \( v = c_n a^m/\sqrt{V} \)

\[ \Prob(z^*_n(n, k) \leq a) \leq (kv)^{n-k} e^{k^2 \sqrt{n-2}/2 \pi k} \quad (2.4) \]

if \( \| \cdot \|_{\infty} \) is used.

**Proof.** Let \( X_1, \ldots, X_k \in \PART(n, k) \) = the set of (unordered) partitions of \( X \) into \( k \) subsets. Let \( n_t = |X_t| \) for \( t = 1, \ldots, k \) and let

\[ Q = \Prob((X_t \subseteq HS(c_t, a) \text{ for some } c_t \in D) \text{ for } t = 1, \ldots, k) \]

\[ = \prod_{t=1}^{k} \Prob(X_t \subseteq HS(c_t, a) \text{ for some } c_t \in D) \]

\[ = \prod_{t=1}^{k} \Prob(F(n_t, a)) \quad \text{as in Lemma 2.2.} \]

By Lemma 2.2

\[ Q \leq v^{n-k} \prod_{t=1}^{k} b^{\sqrt{n_t}} \leq v^{n-k} b^{\sqrt{n_k}}. \]
Now let \( S(X) = \{(X_1, \ldots, X_k) \in \text{PART}(n, k) : r(X_i) \leq a \text{ for } t = 1, \ldots, k\} \) (\( r \) as in Lemma 2.1). We note that \( S(X) = \emptyset \rightarrow z^*_{1}(n, k) > a \). Thus
\[
\text{Prob}(z^*_{1}(n, k) \leq a) \leq \text{Prob}(S(X) \neq \emptyset) \\
\leq E(|S(X)|) \quad \text{(Expectation by the above)} \\
\leq |\text{PART}(n, k)| v^{n-k} b^{\sqrt{n}k} \\
= (k^n/k!) v^{n-k} b^{\sqrt{n}k}
\]
The result now follows after using Stirlings inequalities. 

**Theorem 2.1.** For sequence of problems where \( n \to \infty \) and \( k = nd + O(1/n^3) \) with \( 0 \leq d < 1 \) we have
\[
\text{Prob}(z^*_{1}(n, k) \leq (\alpha_d V/kc_m)^{1/m}) \leq 1/\sqrt{2\pi k} + O(1/n)
\]
(2.5) where \( \alpha_d = (e^{-d}b^{-d})^{1/(1-d)} \).

**Proof.** Simply substitute \((\alpha_d V/kc_m)^{1/m}\) for \( a \) in (2.4).

Thus if \( k \to \infty \) (2.5) provides a lower bound for \( z^*_{1}(n, k) \) with probability tending to 1.

For constant \( k \) we must clearly have \( z^*_{1}(n, k) \geq (V/kc_m)^{1/m} \) – note \( \alpha_0 = 1 \) – else we cannot cover \( D \) with \( k \) hyperspheres of radius \( z^*_{1}(n, k) \). It is straightforward to show that for finite \( k \) we must do this with probability tending to 1.

We continue by computing lower bounds to \( z^*_{1}(n, k) \) for \( \| \cdot \|_\infty \). We base our analysis on a lemma about covering \( D \) with hypercubes. It will be used for sets of the form \( \{x \in R^m : \|x - c\| \leq a\} \).

**Notation.** Let \( a, c \in R^m a \geq 0 \). The hyperoblong is
\[
\text{HO}(c, a) = \{x \in R^m : |x_j - c_j| \leq \frac{1}{2}a_j \text{ for } j = 1, \ldots, m\}.
\]
It’s hypervolume is of course \( a_1a_2 \cdots a_m \).

**Lemma 2.5.** Let \( E(n, c, a) \) be the event that \( n \) points \( X = \{x^{(1)}, \ldots, x^{(n)}\} \) chosen at random from \( D \) lie in \( \text{HO}(c, a) \) and let \( F(n, a) = \bigcup_{c \in D} E(n, c, a) \). Then for \( n \geq 2 \)
\[
P(n, a) = \text{Prob}(F(n, a)) \leq n^m v^{n-1}
\]
where \( v = a_1a_2 \cdots a_m/V \).

**Proof.** Let \( p(n, z) \) be the density function of the random vector \( z \in R^m \) where \( z_1, \ldots, z_m \) are the lengths of the sides of the smallest hyperoblong containing the set \( X \). These lengths are given by
\[
z_i = \max(x^{(1)}_i, \ldots, x^{(n)}_i) - \min(x^{(1)}_i, \ldots, x^{(n)}_i), \]
Thus
\[ P(n, a) = \int_0^{a_1} \cdots \int_0^{a_m} p(n, z) \, dz_m \cdots dz_1 \quad \text{and} \quad p(n, z) = \frac{\partial^n p(n, z)}{\partial z_1 \cdots \partial z_m}. \]

We also have
\[ P(n + 1, a) \leq \int_0^{a_1} \cdots \int_0^{a_m} \frac{p(n, z)}{V} \left( \prod_{i=1}^{m} (2a_i - z_i) \right) \, dz_m \cdots dz_1. \quad (2.6) \]

This is because for given \( z_1 \cdots z_m \) the random point \( x^{(n+1)} \) must lie in a hyperoblong of sides \( (2a_1 - z_1) \cdots (2a_m - z_m) \) in order that \( F(n + 1, a) \) can occur.

Next let \( M = \{1, 2, \ldots, m\} \) and for \( S \subseteq M \) let \( P_S \) denote \( P(n, h_1 \cdots h_m) \) where \( h_i = a_i \) for \( i \in S \) and \( h_i = z_i \) for \( i \notin S \). Let \( d_S = \prod_{i \notin S} dz_i \) and \( a_S = \prod_{i \in S} a_i \). Successive integration of the RHS of (2.6) by parts gives
\[ P(n + 1, a) \leq \left( \sum_{S \subseteq M} a_S \int P_S \, d_S \right) / V. \quad (2.7) \]

Now \( P(2, a) \leq 2^m a_n / V \) and if for some \( n \geq 2 \) and constant \( \alpha \) \( P(n, z) \leq \alpha (\prod_{i=1}^{m} (z_i) / V)^n - 1 \) for \( z \geq 0 \) then from (2.7) we have
\[ P(n + 1, a) \leq \left( \sum_{S \subseteq M} \alpha a_S^n \int \left( \prod_{i \notin S} z_i^{n-1} \, dz_i \right) / V^n \right) \]
\[ = \left( \sum_{S \subseteq M} \alpha a_S^n \left| S \right| / V^n \right), \quad \left| S \right| = M - S \]
\[ = \alpha (1 + 1/n)^n v^n. \]

Thus
\[ P(n + 1, a) \leq \prod_{i=1}^{n} (1 + 1/t)^n v^n = (n + 1)^n v^n. \quad \square \]

**Lemma 2.7.** Let \( n \) points \( X = \{x^{(1)}, \ldots, x^{(n)}\} \) be chosen at random in \( D \). For \( a \geq 0 \) and \( v = a^n / V \)
\[ \text{Prob}(z^*(n, k) \leq \frac{1}{2} a) \leq (kv)^{n-k} e^k (n/k)^3 / \sqrt{2\pi k}. \quad (2.8) \]

**Proof.** Let \( (X_1, \ldots, X_k) \in \text{PART}(n, k) \) and \( |X_i| = n_i \) for \( i = 1, \ldots, k \). Let \( a = (a, \ldots, a) \in R^m \) and let \( Q = \text{Prob}((X_i \subseteq \text{HO}(c_i, a) \text{ for some } c_i \in D) \text{ for } i = 1, \ldots, k) \).

By Lemma 2.4 with \( v = a^n / V \)
\[ Q \leq \prod_{i=1}^{k} n_i^m v^{n-1} \leq (n/k)^m v^{n-k}. \]
It follows as in Lemma 2.3 that
\[
\text{Prob}(z^*_1(n, k) \leq a/2) \leq (k^n/k!) Q.
\]
The result now follows after using Stirling inequalities. \(\square\)

**Theorem 2.2.** For a sequence of problems where \(n \to \infty\) and \(k = nd + O(1/n)\) with \(0 \leq d < 1\) we have
\[
\text{Prob}(z^*_1(n, k) \leq (\frac{1}{2} \alpha_d V/k)^{1/m}) \leq 1/\sqrt{2\pi k} + O(1/n) \tag{2.9}
\]
where \(\alpha_d = (e^{-d} d^{md})^{1/(1-d)}\).

**Proof.** Simply substitute \((\alpha_d V/k)^{1/m}\) for \(a\) in (2.8). \(\square\)

Similar comments to those given after Theorem 2.1 apply. We now describe the calculation of upper bounds and approximate solutions in the case that \(D\) is a hypercube of side \(L\). Let \(\hat{k} = \lfloor k^{1/m} \rfloor\) and divide \(D\) uniformly into \(\hat{k}^m\) hypercubes of side \(L/\hat{k}\) and let \(Y\) consist of the centres of these hypercubes plus \(k - \hat{k}^m\) other points in \(D\).

\(\|\|_{\ell}\). For \(x \in D\) there is a point \(y \in Y\) such that \(\|x - y\|_{\ell} \leq m^{1/2} L/2\hat{k}\) and so for this norm
\[
z^*_1(n, k) \leq m^{1/2} L/2\hat{k}.
\]

\(\|\|_{\infty}\). For \(x \in D\) there is a point \(y \in Y\) such that \(\|x - y\|_{\infty} \leq L/2\hat{k}\) and so for this norm
\[
z^*_1(n, k) \leq L/2\hat{k}.
\]

Notice that if \(k = \hat{k}^m\) this upper bound coincides closely with the lower bound derived after Theorem 2.2 when \(d = 0\).

We now consider approximate solutions. Let \(t > 0\) be an integer which determines the proposed accuracy of the solution. Divide \(D\) uniformly into \(T = t^m\) hypercubes \(H_1, \ldots, H_T\) of side \(L/t\). Let \(C = \{c_1, \ldots, c_T\}\) be the set of centres of these hypercubes. Let
\[
\hat{z}_i = \min(z_1(X, Y) : Y \subseteq C \text{ and } |Y| = k).
\]
This can be computed in \(O(2^T nk)\) time. Let \(Z^*\) minimise \(z_1\). Assume without loss of generality that \(Z^* \subseteq \bigcup_{j=1}^k H_j\). Now for \(x \in D\) and \(y \in H_j\)
\[
\|c_j - x\| \leq \|c_j - y\| + \|y - x\|
\]
and hence
\[
\hat{z}_i - z^*_i \leq \max(\|c_i - y\| : y \in Z^* \cap H_i).
\]
\(\|\|_{\ell}\). Thus \(\hat{z}_i - z^*_i \leq m^{1/2} L/2t\). Now fix \(1 > \varepsilon > 0\) and consider a sequence of problems for which \(k \leq p \log n\) where \(p > 0\). Putting \(t = \lceil m^{1/2}(kc_m)^{1/m}/2\varepsilon \rceil\) we see that \(\hat{z}_i - z^*_i \leq \varepsilon z^*_i\) with probability \(\geq 1 - (2\pi k)^{-1/2}\).
For large $k$ $2^{T} = A^k$ where $A = 2^{(m - 2e_{+}^2/2^e_{-})}$
\[ \leq n^{p \log A} \]
and so the approximation scheme is polynomial when $k$ is restricted in this manner.

In this case $\| z_1 - z_1^* \|_{\infty} \leq L/2t$ and we take $t = \lceil k^{1/m}/\varepsilon \rceil$.

### 3. Analysis of problem 2

We first compute a probabilistic lower bound for problem 2 using $\| \|_c$.

**Lemma 3.1.** Let $n$ points $X = \{x^{(1)}, \ldots, x^{(n)}\}$ be chosen at random in $D$. For $a \geq 0$ and $v = c_m a^m/V$

\[
\text{Prob}(z_1^*(n, k) \leq a) \leq (12/11)(kv)^{n-k} (n^n/k^k (n-k)^{n-k})^{1/2n} \pi^k (n-k).
\]

(3.1)

**Proof.** Let $J = \{j_1, \ldots, j_k\} \subseteq N = \{1, 2, \ldots, n\}$ and let $Y = \{x^{(j_1)}, \ldots, x^{(j_k)}\}$. If $j \in N - J$, then $\text{Prob}(\text{there exists } i(j) \in J \text{ such that } \| x^{(i)} - x^{(i(j))} \|_c \leq a) \leq kv$. Hence

\[
\text{Prob}(\text{for all } j \in N - J \text{ there exists } i(j) \in J \text{ such that } \| x^{(i)} - x^{(i(j))} \|_c \leq a) \leq (kv)^{n-k}
\]

(3.2)

Now there are $\binom{n}{k}$ subsets of size $k$ in $N$ and hence $\text{Prob}(z_1^*(n, k) \leq a) = \text{Prob}(3.2)$ holds for some $J \leq \binom{n}{k}(kv)^{n-k}$. the result now follows after using Stirlings inequalities.

**Theorem 3.1.** For a sequence of problems where $n \to \infty$ and $k = nd + O(1/n)$ with $0 \leq d < 1$ we have

\[
\text{Prob}(z_1^*(n, k) \leq (\alpha_d V/kc_m)^{1/m}) \leq (12/11)\sqrt{n/2\pi k (n-k)} + O(1/n)
\]

(3.3)

where $\alpha_d = (1 - d)d^{d/(1-d)}$.

**Proof.** Use Lemma 3.1. \(\square\)

Thus if $k \to \infty$ (3.3) provides a lower bound for $z_1^*(n, k)$ with probability tending to 1.

For constant $k$ the problem can be solved exactly in $O(n^{k+1})$ time by examining each $k$-subset of $X$.

In the case of $\| \|_c$ a similar proof gives

**Theorem 3.2.** For a sequence of problems where $n \to \infty$ and $k = nd + O(1/n)$ with $0 \leq d < 1$ we have

\[
\text{Prob}(z_1^*(n, k) \leq (\alpha_d V/k)^{1/m}) \leq (12/11)\sqrt{n/2\pi k (n-k)} + O(1/n)
\]

(3.3)

where $\alpha_d = (1 - d)d^{d/(1-d)}$. 
We once again describe the calculation of upper bounds and approximate solutions in the case that $D$ is a hypercube of side $L$. We again divide $D$ uniformly into $\hat{k}^m$ hypercubes of side $L/k$ and this time to produce $Y$ we select one point of $X$ from each hypercube that contains points of $X$ and then make up $Y$ to size $k$ be arbitrary addition of points in $X$ not used so far. This gives

$$z_2^* \leqslant m^{1/2} L/k \quad \text{for} \quad \|\|_e,$$
$$z_2^* \leqslant L/k \quad \text{for} \quad \|\|_e.$$

To obtain approximate solutions we proceed in much the same manner as in Section 2. We choose $t > 0$ as before and divide $D$ into $H_1, \ldots, H_T$. For each $J \subseteq SJ = \{J \subseteq \{1, \ldots, T\} : |J| = k\}$ we proceed as follows: for each $j \in J$ such that $H_j \cap X \neq \emptyset$ choose $(x^{(j)}) \in H_j \cap X$. This produces $k \leqslant k$ points to which we arbitrarily add $k - k_1$ other points from $X$ to form a set $Y(J)$. Then let $\tilde{z}_2 = \min(z_2(X, Y(J)) : J \subseteq SJ)$ which can be computed in $O(2^{Tn})$ time.

Now let $Y^*$ minimize $z_3$ and assume without loss of generality that $Y^* \subseteq \bigcup_{j=1}^k H_j$. A use of the triangular inequality as in Section 2 shows that

$$\tilde{z}_2 - z_2^* \leqslant L_t \quad \text{where} \quad L_t = \max(\|x - y\| : x, y \in H_t).$$

Assuming $k \leqslant d \log n$ and given $\epsilon > 0$ and taking

$$t = \left\lfloor m^{1/2}(k\epsilon m)^{1/m}/\epsilon \right\rfloor \quad \text{for} \quad \|\|_e,$$
$$t = \left\lfloor 2k^{1/m}/\epsilon \right\rfloor \quad \text{for} \quad \|\|_x,$$

we have $\tilde{z}_2 - z_2^* \leqslant \epsilon z_2^*$ with high probability and the time taken is polynomial in $n$.

4. Analysis of problem 3

Our lower bounds for $\|\|_e$ are based on

**Lemma 4.1.** Let $X \subseteq R^m$ be a finite set and suppose that $x, y \in X$ implies $\|x - y\|_e \leqslant a$. Then $r = r(X) \leqslant a(m/2(m + 1))^{1/2}$ where $r$ is the radius of the smallest hypersphere containing $X$.

**Proof.** Let $c = c(X)$ be the centre of this hypersphere and as in Lemma 2.1 $c = \sum_{i=1}^d \lambda_i y_i$ where $\|y_i - c\|_e = r$. We can assume by Caratheodory’s theorem that $d \leqslant m + 1$. If $z_i = (y_i - c)/r$ for $1 \leqslant i \leqslant d$ then

$$0 = \left\|\sum_{i=1}^d \lambda_i z_i\right\|_e^2 = \sum \lambda_i^2 + 2 \sum \lambda_i \lambda_j z_i \cdot z_j.$$

We show that there exists $k, l$ such that $z_k \cdot z_l \leqslant -1/(d - 1)$. (If $d = 1$ then $X = \{c\}$ and the result is trivial.) For if not we have

$$0 > \sum \lambda_i^2 - (2/(d - 1)) \sum \lambda_i \lambda_j = (\sum (\lambda_i - \lambda_j)^2)/(d - 1) \geqslant 0.$$
Thus
\begin{align*}
a^2 &\geq \|y_k - y_l\|^2 \\
&= r^2 \|z_k - z_l\|^2 \\
&= r^2(z_k^2 + z_l^2 - 2z_k \cdot z_l) \\
&\geq r^2(2 + 2/(d-1)) \geq r^2(2 + 2/m). \quad \square
\end{align*}

Using this result in conjunction with Theorem 2.1 gives

**Theorem 4.1.** For a sequence of problems where \( n \to \infty \) and \( k = nd + O(1/n) \) with \( 0 \leq d < 1 \) we have

\[
\text{Prob}(\hat{z}_3^*(n, k) \leq \alpha_d (V/kc_m)^{1/m}) \leq 1/\sqrt{2\pi k} + O(1/n)
\]

where \( \alpha_d = (2(m+1)/m)^{1/2}(e^{-d} b^{-\sqrt{d}})^{1/m(1-d)} \).

The result for \( \| \|_k \) depends on the fact that if X \( \subseteq \mathbb{R}^m \) is such that \( x, y \in X \) implies \( \|x - y\|_k \leq a \) then X can be contained in a hypercube of side \( a \). This gives using Theorem 2.2.

**Theorem 4.2.** For a sequence of problems for which \( n \to \infty \) and \( k/n \to d < 1 \) we have

\[
\text{Prob}(\hat{z}_3^*(n, k) \leq (\alpha_d V/k)^{1/m}) \leq 1/\sqrt{2\pi k} + O(1/n)
\]

where \( \alpha_d = (e^{-d} d^{md})^{1/(1-d)} \).

Once again assuming that D is a hypercube of side \( L \) we obtain upper bounds by dividing D into \( \tilde{k} = \tilde{k}^m \) hypercubes of side \( L/\tilde{k} \). Let these hypercubes be \( H_1, \ldots, H_{\tilde{k}} \). We then partition X into \( X \cap H_1, \ldots, X \cap H_{\tilde{k}} \) plus \( k - \tilde{k} \) empty sets. If there are points of X on the boundaries of several hypercubes we assign these points arbitrarily to one of them. This partition gives

\[
z_3^* \leq \frac{1/2}{L/\tilde{k}} \quad \text{for} \quad \| \|_k
\]
\[
z_3^* \leq L/\tilde{k} \quad \text{for} \quad \| \|_x.
\]

To obtain approximate solutions we again choose \( t > 0 \) and divide D into \( T = t^m \) hypercubes \( H_1, \ldots, H_T \). For \( J \subseteq \{1, 2, \ldots, T\} \) let \( H_J = \bigcup_{i \in J} H_i \) and let \( P(T) \) the set of partitions of \( \{1, \ldots, T\} \) into \( k \) subsets. For \( (J_1, \ldots, J_k) \in P(T) \) let

\[
Z_3(J_1, \ldots, J_k) = \max(\max(\|x - y\|: x, y \in H_{i} \cap X): i = 1, \ldots, k)
\]

and let \( \hat{z}_3 = \min(Z_3(J_1, \ldots, J_k): (J_1, \ldots, J_k) \in P(T)) \). \( \hat{z}_3 \) can be computed in \( O((k^T/k)!n^2) \) time. Now let \( (\hat{x}_1^*, \ldots, \hat{x}_k^*) \) be the optimal partition for \( z_3 \). The partitions generated in computing \( \hat{z}_3 \) are all those that satisfy

\[
X_i \cap H_j \neq \emptyset \text{ for some } i, \text{ } r \text{ implies } X_i \cap H_r = \emptyset \text{ for } i \neq j.
\]
If \(X^*_1, \ldots, X^*_k\) does not satisfy (4.3) then we can find \((\tilde{X}_1, \ldots, \tilde{X}_k)\) satisfying (4.3) and
\[
z_3(\tilde{X}_1, \ldots, \tilde{X}_k) \leq z_3(X^*_1, \ldots, X^*_k) + 2L_r.
\]
(\(\tilde{X}_1, \ldots, \tilde{X}_k\)) is obtained by starting with \((X^*_1, \ldots, X^*_k)\) and while there are \(r, i, j_1, \ldots, j_p\) contravening (4.3) amending the current partition \((X_1, \ldots, X_k)\) by \(X_i := X_i \cup \bigcup_{r=1}^{p} (X_{j_r} \cap H_r)\) and \(X_{j_r} := X_{j_r} - H_r\) for \(s = 1, \ldots, p\). We observe that throughout the above process
\[
x \in X_i \implies \text{there exists } y, r \text{ such that } y \in X^*_i \text{ and } x, y \in H_r
\]
(either \(y = x\) or prior to some change of partition \(x \in X_{j_r} \cap H_r\) and it is then moved to \(X_i\) while \(y \in X_i \cap H_r\) is never moved).

Let \(z_3(\tilde{X}_1, \ldots, \tilde{X}_k) = \|\tilde{x} - \tilde{y}\|\) where \(\tilde{x}, \tilde{y} \in \tilde{X}_p\). From (4.5) there exist \(x^*, y^*, r, s\) such that \(x^*, y^* \in X^*_p\), \(\tilde{x}, x^* \in H_r\) and \(\tilde{y}, y^* \in H_s\). Thus
\[
\|\tilde{x} - \tilde{y}\| \leq \|\tilde{x} - x^*\| + \|x^* - y^*\| + \|y^* - \tilde{y}\|
\]
\[
\leq L_r + z_3^* + L_r
\]
and thus \(\tilde{z}_3 \leq z_3^* + 2L_r\).

Let \(t = m^{1/2}/t\). Now fix \(1 < \epsilon < 0\) and take
\[t = \lceil m(k/c)^{1/m}/((m+1)/2)^{1/2} \epsilon \rceil.
\]
This gives \(z_3 - z_3^* \leq \epsilon z_3^*\) with high probability. The dominant term in \(kT/k!\) is, using Stirling's formulae \(A^k\) for some constant \(A\) and for polynomial time we again assume \(k \leq d \log n\).

5. Analysis of problem 4

We first compute a probabilistic lower bound to problem 4 using \(\|\|_c\|\).

**Lemma 5.1.** Let \(X = \{x^{(1)}, \ldots, x^{(n)}\}\) be chosen at random in \(D\). Let \(z(X) = \sum_i \sum_j \|x^{(i)} - x^{(j)}\|_c\). Then
\[
\text{Prob}(z(X) \leq a) \leq n (m! c_{m}(a/n)^{m}/V)^{n-1}/(m(n-1))!.
\]

**Proof.** We first consider the following: \(c\) is an arbitrary point of \(D\) and \(Y = \{y^{(1)}, \ldots, y^{(n)}\}\) are chosen at random in \(D\). Let \(d_i = \|c - y^{(i)}\|_c\) and \(d(c, Y) = \sum_i d_i\). Then
\[
\text{Prob}(b_i \leq d_i \leq b_i + \delta b_i) \leq mc_{m}b_i^{n-1} \delta b_i(1 + O(\delta b_i))/V
\]
as \(mc_{m}b_i^{n-1} \delta b_i\) is the approximate hypervolume of a "thin hyperannulus" of
radius $b_t$ and thickness $\delta b_t$. Consequently

$$\text{Prob}(d(c, Y) \leq a) \leq \int_{b_0 = 0}^{b_1 - a} \int_{b_2 = a - b_1}^{b_2} \cdots \int_{b_n = a - \sum_{i=1}^{n} b_i}^{b_n} \prod_i (mc_m b_i^{n-1} \, db_i/V)$$

$$= (m! c_m a^n / V)^n / (qm)!$$

which is easily proved by induction. Putting $X_j = X/\{x^{(j)}\}$ we see that

$$\text{Prob}(d(x^{(j)}, X_j) \leq a) \leq (m! c_m a^n / V)^{n-1} / ((n-1)m)!.$$  \hfill (5.2)

Now

$$\text{Prob} \left( \sum_{j=1}^n d(x^{(j)}, X_j) = na \right) \leq \text{Prob}(d(x^{(j)}, X_j) \leq a \text{ for at least one } j = 1, \ldots, n)$$

$$\leq n \text{ Prob}(d(x^{(1)}, X_1) \leq a).$$ \hfill (5.3)

We obtain (5.1) by replacing $a$ in (5.3) by $a/n$ and using (5.2). ∎

**Lemma 5.2.** Let $X = \{x^{(1)}, \ldots, x^{(n)}\}$ be chosen at random in $D$. For $a \geq 0$ and assuming $n \geq 2k$, then

$$\text{Prob}(z_{m_k}^p(n, k) \leq a) \leq A/B\hfill \hfill (5.4)$$

where

$$A = e^{m(n-1)+k(2m+1)n-2(m+1)k}n^k(m! c_m a^n / V)^{n-k},$$

$$B = (m(n-k)^2)^{m(n-k)}(2\pi mk(n-k))^{1/2}.$$

**Proof.** Let $(X_1, \ldots, X_k) \in \text{PART}(n, k)$ and let $n_t = |X_t|$ for $t = 1, \ldots, k$ and let $Q = \text{Prob}(z(X_t) \leq a \text{ for } t = 1, \ldots, k)$ where $z$ is as defined in the statement of Lemma 4.1. From this lemma

$$Q \leq \prod_{i=1}^k n_t (m! c_m a^n / V)^{n_t-1} / n_t^m (n_t-1)!$$

$$\leq (n/k)^k (m! c_m a^n / V)^{n-k} / P$$

where

$$P = \prod_{n_t \geq 2} (mn_t(n_t-1))^{m(n_t-1)}(2\pi m(n_t-1))^{1/2}e^{-m(n_t-1)}$$

$$\geq m^{m(n-k)^2}((n-k)/k)^{2m(n-k)}(2\pi m(n-k))^{1/2}e^{-m(n-k)}$$

where $n \geq 2k$ is used in one of the reductions.

As there are at most $k^n/k!$ partitions we have our result in the usual way after using Stirling's inequalities. ∎

**Theorem 5.1.** For a sequence of problems where $n \to \infty$ and $k = nd + O(1/n)$, and
(i) $0 < d \leq \frac{1}{2}$:

$$\text{Prob}(z^*_d(n, k) \leq \alpha_d(V/kc_m)^{1/m}) \leq 1/\sqrt{2\pi mk(n-k) + O(1/n)}$$

where $\alpha_d = (1-d)/d)^2 m((d^d e^{-m-d})^{1/(1-d)/m})^{1/m}.$

(ii) $d = 0$:

$$\text{Prob}(z^*_d(n, k) \leq \beta_m(n/k)^2 n^{-k/n}(V/kc_m)^{1/m}) \leq 1/\sqrt{2\pi mk(n-k) + O(1/n)}$$

where $\beta_m = m/e(m!)^{1/m}.$

Proof. Use Lemma 5.2. □

For $\| \cdot \|_\infty$ we have

**Theorem 5.2.** For a sequence of problems where $n \to \infty$ and $k = nd + O(1/n)$ and

(i) $0 < d \leq \frac{1}{2}$:

$$\text{Prob}(z^*_d(n, k) \leq \alpha_d(V/k)^{1/m/2}) \leq 1/\sqrt{2\pi mk(n-k) + O(1/n)}.$$

(ii) $d = 0$:

$$\text{Prob}(z^*_d(n, k) \leq \beta_m(n/k)^2 n^{-k/n}(V/k)^{1/m/2}) \leq 1/\sqrt{2\pi mk(n-k) + O(1/n)}$$

where $\alpha_d, \beta_m$ are as in Theorem 5.1.

Once again assuming that $D$ is a hypercube of side $L$ we obtain an upper bound by using the partition defined in Section 4. This gives

$$z^*_4 \leq m^{1/2} L n^2/\hat{k} \quad \text{for} \quad \| \cdot \|_1,$$

$$z^*_4 \leq L n^2/\hat{k} \quad \text{for} \quad \| \cdot \|_\infty.$$

Note that these upper bounds are larger than the lower bounds by a factor of order of magnitude $k^2$. This can be explained by the possibility that all the points of $X$ lie in the same small hypercube. To obtain approximate solutions we proceed as in Section 4 to compute

$$\hat{z}_4 = \min(Z_4(J_1, \ldots, J_k) : (J_1, \ldots, J_k) \in P(T))$$

where

$$Z_4(J_1, \ldots, J_k) = \max\left(\sum_{x,y \in H_i \cap X} \|x-y\| : i = 1, \ldots, k\right)$$

assuming the usual uniform division of $D$ into $t^m$ hypercubes. Once again let $(X^*_1, \ldots, X^*_k)$ minimise $z_4$. If this partition satisfies (4.3) then $\hat{z}_4 = z^*_4$ otherwise we can compute from it a partition $(\hat{X}_1, \ldots, \hat{X}_k)$ satisfying (4.3) and

$$z_4(\hat{X}_1, \ldots, \hat{X}_k) \leq z^*_4 + n^2 L_4.$$  \hspace{1cm} (5.5)

We start from $(X^*_1, \ldots, X^*_k)$ and a general stage of the construction suppose we have the partition $(X_1, \ldots, X_k)$ and for some $r \not\in I \not\geq 2$ where $I = \{i : X_i \cap H_i \neq \emptyset\}$. Let $e$ be the centre of $H$ and for $i \in I$ let $d_i = \sum_{x \in X_i} \|x-e\|$ and let $d_p = \min(d_i)$. We
amend the current partition as follows:

\[ X_p := X_p \cup (H_i \cap X), \]

\[ X_i := X_i - H_i, \quad i \in I - \{p\}. \]

The change \( \Delta \) in the value of \( z_4 \) is

\[ \sum_{i \in I - \{p\}} \sum_{x \in H_i \cap X_i} \left( \sum_{y \in X_p} \|x - y\| - \sum_{y \in X_i - H_i} \|x - y\| \right). \]

Now

\[ \sum_{x \in H_i \cap X_i} \sum_{y \in X_p} \|x - y\| \leq n_i (d_p + |X_p| L_d/2) \]

where \( n_i = |H_i \cap X_i| \) and

\[ \sum_{x \in H_i \cap X_i} \sum_{y \in X_i - H_i} \|x - y\| \geq n_i (d_i - |X_i| L_d/2) \]

and so

\[ \Delta \leq \sum_{i \in I - \{p\}} n_i (d_p - d_i + (|X_p| + |X_i|) L_d/2) \]

\[ \leq \sum_{i \in I - \{p\}} n_i n L_d, \]

Continuing this process until \((4.3)\) is satisfied we find that the total change is \( \leq n^2 L_d \).

\[ \|z_4 - z_4^*\| \leq 2m^{1/2} n^2 L_d/\epsilon. \]

Now fix \( 1 < \epsilon < 0 \) and take

\[ t = \left[ 2m^{1/2} k^2 (kC_m)^{1/m} / \epsilon \beta m \right]. \]

This gives \( \hat{z}_4 - z_4^* \leq \epsilon z_4^* \) with high probability. The time for computing \( z_4 \) is \( O(n^2 k^{T+1}/k!) \) and the dominant term in \( k^T/k! \) is one of \( k^{k^2 + k/m} \) and to get a polynomial time algorithm we assume \( k \approx (b \log n / \log \log n)^{1/(2+1/m)} \).

\[ \|z_4 - z_4^*\| \leq \epsilon z_4^*. \]

Here we take \( t = \left[ 2k^{2+1/m} / \epsilon \beta m \right]. \)

Acknowledgement

I am grateful to Colin McDiarmid for his valuable comments.

References

