A Las Vegas approximation algorithm for metric 1-median selection

Ching-Lueh Chang *†

February 28, 2017

Abstract

Given an $n$-point metric space, consider the problem of finding a point with the minimum sum of distances to all points. We show that this problem has a randomized algorithm that always outputs a $(2+\epsilon)$-approximate solution in an expected $O(n/\epsilon^2)$ time for each constant $\epsilon > 0$. Inheriting Indyk’s [9] algorithm, our algorithm outputs a $(1+\epsilon)$-approximate 1-median in $O(n/\epsilon^2)$ time with probability $\Omega(1)$.

1 Introduction

A metric space is a nonempty set $M$ endowed with a metric, i.e., a function $d: M \times M \to [0, \infty)$ such that

- $d(x, y) = 0$ if and only if $x = y$ (identity of indiscernibles),
- $d(x, y) = d(y, x)$ (symmetry), and
- $d(x, y) + d(y, z) \geq d(x, z)$ (triangle inequality)

for all $x, y, z \in M$ [13].

For all $n \in \mathbb{Z}^+$, define $[n] \equiv \{1, 2, \ldots, n\}$. Given $n \in \mathbb{Z}^+$ and oracle access to a metric $d: [n] \times [n] \to [0, \infty)$, METRIC 1-MEDIAN asks for $\arg\min_{y \in [n]} \sum_{x \in [n]} d(y, x)$, breaking ties arbitrarily. It generalizes the classical median selection on the real line and has a brute-force $\Theta(n^2)$-time algorithm. More generally, METRIC $k$-MEDIAN asks for $c_1, c_2, \ldots, c_k \in [n]$ minimizing $\sum_{x \in [n]} \min_{i=1}^k d(x, c_i)$. Because $d(\cdot, \cdot)$ defines $\binom{n}{2} = \Theta(n^2)$ nonzero distances, only $o(n^2)$-time algorithms are said

*Department of Computer Science and Engineering, Yuan Ze University, Taoyuan, Taiwan. Email: clchang@saturn.yzu.edu.tw
†Supported in part by the Ministry of Science and Technology of Taiwan under grant 105-2221-E-155-047-.
to run in sublinear time [8]. For all $\alpha \geq 1$, an $\alpha$-approximate 1-median is a point $p \in [n]$ satisfying
\[
\sum_{x \in [n]} d(p, x) \leq \alpha \cdot \min_{y \in [n]} \sum_{x \in [n]} d(y, x).
\]

For all $\epsilon > 0$, METRIC 1-MEDIAN has a Monte Carlo $(1 + \epsilon)$-approximation $O(n/\epsilon^2)$-time algorithm [8, 9]. Guha et al. [7] show that METRIC $k$-MEDIAN has a Monte Carlo, $O(\exp(O(1/\epsilon)))$-approximation, $O(nk \log n)$-time, $O(n^k)$-space and one-pass algorithm for all small $k$ as well as a deterministic, $O(\exp(O(1/\epsilon)))$-approximation, $O(n^{1+\epsilon})$-time, $O(n^k)$-space and one-pass algorithm. Given $n$ points in $\mathbb{R}^D$ with $D \geq 1$, the Monte Carlo algorithms of Kumar et al. [10] find a $(1 + \epsilon)$-approximate 1-median in $O(D \exp(1/\epsilon^{O(1)}))$ time and a $(1 + \epsilon)$-approximate solution to METRIC $k$-MEDIAN in $O(Dn \cdot \exp((k/\epsilon)^{O(1)}))$ time. All randomized $O(1)$-approximation algorithms for METRIC $k$-MEDIAN take $\Omega(nk)$ time [7, 11]. Chang [2] shows that METRIC 1-MEDIAN has a deterministic, $(2h)$-approximation, $O(hn^{1+1/h})$-time and nonadaptive algorithm for all constants $h \in \mathbb{Z}^+ \setminus \{1\}$, generalizing the results of Chang [1] and Wu [15]. On the other hand, he disproves the existence of deterministic $(2h - \epsilon)$-approximation $O(n^{1+1/(h-1)/h})$-time algorithms for all constants $h \in \mathbb{Z}^+ \setminus \{1\}$ and $\epsilon > 0$ [3, 4].

In social network analysis, the closeness centrality of a point $v$ is the reciprocal of the average distance from $v$ to all points [14]. So METRIC 1-MEDIAN asks for a point with the maximum closeness centrality. Given oracle access to a graph metric, the Monte-Carlo algorithms of Goldreich and Ron [6] and Eppstein and Wang [5] estimate the closeness centrality of a given point and those of all points, respectively.

All known sublinear-time algorithms for METRIC 1-MEDIAN are either deterministic or Monte Carlo, the latter having a positive probability of failure. For example, Indyk’s Monte Carlo $(1 + \epsilon)$-approximation algorithm outputs with a positive probability a solution without approximation guarantees. In contrast, we show that METRIC 1-MEDIAN has a randomized algorithm that always outputs a $(2 + \epsilon)$-approximate solution in expected $O(n/\epsilon^2)$ time for all constants $\epsilon > 0$. So, excluding the known deterministic algorithms (which are Las Vegas only in the degenerate sense), this paper gives the first Las Vegas approximation algorithm for METRIC 1-MEDIAN with an expected sublinear running time. Note that deterministic sublinear-time algorithms for METRIC 1-MEDIAN can be 4-approximate but not $(4 - \epsilon)$-approximate for any constant $\epsilon > 0$ [1, 4]. So our approximation ratio of $2 + \epsilon$ beats that of any deterministic sublinear-time algorithm. Inheriting Indyk’s algorithm, our algorithm outputs a $(1 + \epsilon)$-approximate 1-median in $O(n/\epsilon^2)$ time with probability $\Omega(1)$ for all constants $\epsilon > 0$.

Below is our high-level and inaccurate sketch of proof, where $\epsilon, \delta > 0$ are small constants:

(i) Run Indyk’s algorithm to find a probably $(1 + \epsilon/10^{10})$-approximate 1-median, $z$. Then let $r = \sum_{x \in [n]} d(z, x)/n$ be the average distance from $z$ to all points.
(ii) For all $R > 0$, denote by $B(z, R)$ the open ball with center $z$ and radius $R$. Use the triangle inequality (with details omitted here) to show $z$ to be a solution no worse than the points in $[n] \setminus B(z, 8r)$, i.e.,

$$\sum_{x \in [n]} d(z, x) \leq \inf_{y \in [n] \setminus B(z, 8r)} \sum_{x \in [n]} d(y, x). \quad (1)$$

(iii) Take a uniformly random bijection $\pi: |B(z, \delta nr)| \to B(z, \delta nr)$. Then observe that

$$\min_{y \in B(z, 8r)} \sum_{x \in B(z, \delta nr)} d(y, x) \geq \min_{y \in B(z, 8r)} \sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} (d(y, \pi(2i-1)) + d(y, \pi(2i))) \quad (2)$$

$$\geq \sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i-1), \pi(2i)), \quad (3)$$

where the first (resp., second) inequality follows from the injectivity of $\pi$ (resp., the triangle inequality).

(iv) Assume $B(z, \delta nr) = [n]$ for simplicity. So by inequalities (1)–(3), if the following inequality holds, then it serves as a witness that $z$ is $(2 + \epsilon)$-approximate:

$$\sum_{x \in B(z, \delta nr)} d(z, x) \leq (2 + \epsilon) \cdot \sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i-1), \pi(2i)). \quad (4)$$

To guarantee outputting a $(2 + \epsilon)$-approximate 1-median, output $z$ only when inequality (4) holds. Restart from item (i) whenever inequality (4) is false.

More details of item (iv) follow: For a 1-median $z'$ of $B(z, \delta nr)$, it will be easy to show

$$\sum_{x \in B(z, \delta nr)} d(z', x) \leq (2 + o(1)) \cdot E \left[ \sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i-1), \pi(2i)) \right]. \quad (5)$$

When $z$ in item (i) is indeed $(1 + \epsilon/10^{10})$-approximate,

$$\sum_{x \in [n]} d(z, x) \leq \left(1 + \frac{\epsilon}{10^{10}}\right) \cdot \sum_{x \in [n]} d(z', x). \quad (6)$$

---

1 Assuming $B(z, \delta nr) = [n]$, inequalities (2)–(4) imply $\sum_{x \in [n]} d(z, x) \leq (2 + \epsilon) \cdot \sum_{x \in [n]} d(y, x)$ for all $y \in B(z, 8r)$. Furthermore, $\sum_{x \in [n]} d(z, x) \leq \sum_{x \in [n]} d(y, x)$ for all $y \in [n] \setminus B(z, 8r)$ by inequality (1).

2 Though not directly stated in later sections, this is a consequence of Lemmas 7 and 12 in Sec. 4.
Assuming $B(z, \delta nr) = [n]$, inequalities (5)–(6) make inequality (4) hold with high probability as long as $\sum_i \lfloor |B(z, \delta nr)|/2 \rfloor d(\pi(2i - 1), \pi(2i))$ is highly concentrated around its expectation. The need for such concentration is why we restrict the radius of the codomain of $\pi$ to be $\delta nr$ in item (iii)—Large distances ruin concentration bounds. To accommodate for the points in $[n] \setminus B(z, \delta nr)$, our witness for the approximation ratio of $z$ actually differs slightly from inequality (4), unlike in item (iv).3

2 Definitions and preliminaries

For a metric space $([n], d)$, $x \in [n]$ and $R > 0$, define

$$B(x, R) \equiv \{y \in [n] \mid d(x, y) < R\}$$

to be the open ball with center $x$ and radius $R$. For brevity,

$$B^2(x, R) \equiv B(x, R) \times B(x, R).$$

The pairs in $B^2(x, R)$ are ordered.

An algorithm $A$ with oracle access to $d: [n] \times [n] \to [0, \infty)$ is denoted by $A^d$ and may query $d$ on any $(x, y) \in [n] \times [n]$ for $d(x, y)$. In this paper, all Landau symbols (such as $O(\cdot)$, $o(\cdot)$, $\Theta(\cdot)$ and $\Omega(\cdot))$ are w.r.t. $n$. The following result is due to Indyk.

**Fact 1** ([8, 9]). For all $\epsilon > 0$, metric 1-median has a Monte Carlo $(1 + \epsilon)$-approximation $O(n/\epsilon^2)$-time algorithm with a failure probability of at most $1/e$.

Henceforth, denote Indyk’s algorithm in Fact 1 by **Indyk median**. It is given $n \in \mathbb{Z}^+$, $\epsilon > 0$ and oracle access to a metric $d: [n] \times [n] \to [0, \infty)$. By convention, denote the expected value and the variance of a random variable $X$ by $E[X]$ and $\text{var}(X)$, respectively.

**Chebyshev’s inequality** ([12]). Let $X$ be a random variable with a finite expected value and a finite nonzero variance. Then for all $k \geq 1$,

$$\Pr \left[ |X - E[X]| \geq k\sqrt{\text{var}(X)} \right] \leq \frac{1}{k^2}.$$

3 Algorithm and approximation ratio

Throughout this paper, take any small constant $\epsilon > 0$, e.g., $\epsilon = 10^{-100}$. By line 1 of **Las Vegas median** in Fig. 1, $\delta > 0$ is likewise a small constant. The following lemma implies that $z$ in line 3 of **Las Vegas median** is a solution (to **metric 1-median**) no worse than those in $[n] \setminus B(z, 8r)$, where $r$ is as in line 4.

---

3Our witness for the approximation ratio of $z$ is as in line 6 of **Las Vegas median** in Fig. 1.
1: Find $\delta > 0$ such that $2 + \epsilon = 2/(1 - 100\sqrt{\delta})$;
2: while true do
3: $z \leftarrow$ Indyk median$^d(n, \epsilon/10^{10})$;
4: $r \leftarrow \sum_{x \in [n]} d(z, x)/n$;
5: Pick a uniformly random bijection $\pi: \lfloor |B(z, \delta nr)| \rfloor \rightarrow B(z, \delta nr)$;
6: if $\sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i - 1), \pi(2i)) + \sum_{x \in [n] \setminus B(z, \delta nr)} (d(z, x) - 8r) \geq (1 - 100\sqrt{\delta})nr/2$ then
7: \textbf{return} $z$;
8: \textbf{end if}
9: \textbf{end while}

Figure 1: Algorithm \textbf{Las Vegas median} with oracle access to a metric $d: [n] \times [n] \rightarrow [0, \infty)$ and with inputs $n \in \mathbb{Z}^+$ and a small constant $\epsilon > 0$

Lemma 2. \textit{In each iteration of the while loop of Las Vegas median,}

$$\inf_{y \in [n] \setminus B(z, 8r)} \sum_{x \in [n]} d(y, x) \geq 7 \cdot \sum_{x \in [n]} d(z, x).$$

\textit{Proof.} For each $y \in [n] \setminus B(z, 8r),$

$$\sum_{x \in [n]} d(y, x) \geq \sum_{x \in [n]} (d(y, z) - d(z, x)) \geq \sum_{x \in [n]} (8r - d(z, x)) = 8nr - \sum_{x \in [n]} d(z, x) = 7 \sum_{x \in [n]} d(z, x),$$

where the first inequality follows from the triangle inequality, the second follows from $y \notin B(z, 8r)$ and the last equality follows from line 4 of \textbf{Las Vegas median}. \hfill $\square$

Lemma 3. \textit{When line 7 of Las Vegas median is run,}

$$\min_{y \in B(z, 8r)} \sum_{x \in [n]} d(y, x) \geq \frac{1 - 100\sqrt{\delta}}{2} \cdot \sum_{x \in [n]} d(z, x).$$

\textit{Proof.} Pick any $y \in B(z, 8r)$. We have

$$\sum_{x \in B(z, \delta nr)} d(y, x) \geq \sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} (d(y, \pi(2i - 1)) + d(y, \pi(2i))) \quad (7)$$

$$\geq \sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i - 1), \pi(2i)).$$
where the first and the second inequalities follow from the injectivity of $\pi$ in line 5 of Las Vegas median and the triangle inequality, respectively. Furthermore,

\[
\sum_{x \in [n] \setminus B(z, \delta nr)} d(y, x) \geq \sum_{x \in [n] \setminus B(z, \delta nr)} (d(z, x) - d(y, z)) \geq \sum_{x \in [n] \setminus B(z, \delta nr)} (d(z, x) - 8r),
\]

(8)

where the first and the second inequalities follow from the triangle inequality and $y \in B(z, 8r)$, respectively. Summing up inequalities (7)–(8),

\[
\sum_{x \in [n]} d(y, x) \geq \left\lfloor \frac{|B(z, \delta nr)|}{2} \right\rfloor \frac{1-100\sqrt{\delta}}{2} \cdot nr
\]

This and lines 6–7 of Las Vegas median imply

\[
\sum_{x \in [n]} d(y, x) \geq \frac{1-100\sqrt{\delta}}{2} \cdot nr
\]

when line 7 is run. Finally, $nr = \sum_{x \in [n]} d(z, x)$ by line 4.

Lemmas 2–3 and line 1 of Las Vegas median yield the following.

**Lemma 4.** When line 7 of Las Vegas median is run,

\[
(2 + \epsilon) \cdot \min_{y \in [n]} \sum_{x \in [n]} d(y, x) \geq \sum_{x \in [n]} d(z, x),
\]

i.e., $z$ is a $(2 + \epsilon)$-approximate 1-median.

By Lemma 4, Las Vegas median outputs a $(2 + \epsilon)$-approximate 1-median at termination.

## 4 Probability of termination in any iteration

This section analyzes the probability of running line 7 in any particular iteration of the while loop of Las Vegas median. The following lemma uses an easy averaging argument.

**Lemma 5.**

\[
| [n] \setminus B(z, \delta nr) | \leq \frac{1}{\delta}
\]

and, therefore,

\[
| B(z, \delta nr) | \geq n - \frac{1}{\delta} = (1 - o(1)) n.
\]

\footnote{Note that $\pi(1), \pi(2), \ldots, \pi(2 |B(z, \delta nr)|/2)$ are distinct elements of $B(z, \delta nr)$.}
Proof. Clearly,

$$\sum_{x \in [n]} d(z, x) \geq \sum_{x \in [n] \setminus B(z, \delta nr)} d(z, x) \geq \sum_{x \in [n] \setminus B(z, \delta nr)} \delta nr = |[n] \setminus B(z, \delta nr)| \cdot \delta nr.$$ 

Then use line 4 of Las Vegas median. \(\square\)

Henceforth, assume \(n \geq 1/\delta + 4\) without loss of generality; otherwise, find a 1-median by brute force. So \(|B(z, \delta nr)| \geq 4\) by Lemma 5. Define

$$r' \equiv \frac{1}{|B(z, \delta nr)|^2} \cdot \sum_{u, v \in B(z, \delta nr)} d(u, v) \tag{9}$$

to be the average distance in \(B(z, \delta nr)\).

Lemma 6. \(r' \leq 2r\).

Proof. By equation (9) and the triangle inequality,

$$r' \leq \frac{1}{|B(z, \delta nr)|^2} \cdot \sum_{u, v \in B(z, \delta nr)} (d(z, u) + d(z, v)) \tag{10}$$

$$= \frac{1}{|B(z, \delta nr)|^2} \cdot |B(z, \delta nr)| \cdot \left( \sum_{u \in B(z, \delta nr)} d(z, u) + \sum_{v \in B(z, \delta nr)} d(z, v) \right)$$

$$= \frac{2}{|B(z, \delta nr)|} \cdot \sum_{u \in B(z, \delta nr)} d(z, u).$$

Obviously, the average distance from \(z\) to the points in \(B(z, \delta nr)\) is at most that from \(z\) to all points, i.e.,

$$\frac{1}{|B(z, \delta nr)|} \cdot \sum_{u \in B(z, \delta nr)} d(z, u) \leq \frac{1}{n} \cdot \sum_{u \in [n]} d(z, u). \tag{11}$$

Inequalities (10)–(11) and line 4 of Las Vegas median complete the proof. \(\square\)

To analyze the probability that the condition in line 6 of Las Vegas median holds, we shall derive a concentration bound for

$$\sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i - 1), \pi(2i)),$$

whose expected value and variance are examined in the next four lemmas.

Lemma 7. With expectations taken over \(\pi\),

$$E \left[ \sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i - 1), \pi(2i)) \right] = \frac{1}{2} \cdot (1 \pm o(1)) nr'. \tag{12}$$
Proof. For each \( i \in \lfloor |B(z, \delta_{nr})|/2 \rfloor \), \( \{\pi(2i - 1), \pi(2i)\} \) is a uniformly random size-2 subset of \( B(z, \delta_{nr}) \) by line 5 of Las Vegas median. Therefore,

\[
E \left[ d(\pi(2i - 1), \pi(2i)) \right] = \frac{1}{|B(z, \delta_{nr})| \cdot (|B(z, \delta_{nr})| - 1)} \sum_{\text{distinct } u, v \in B(z, \delta_{nr})} d(u, v) \tag{13}
\]

\[
= \frac{1}{|B(z, \delta_{nr})| \cdot (|B(z, \delta_{nr})| - 1)} \sum_{u, v \in B(z, \delta_{nr})} d(u, v)
\]

\[
= (1 + o(1)) r',
\]

where the second (resp., last) equality follows from the identity of indiscernibles (resp., equation (9) and Lemma 5). Finally, use equations (13)–(14), the linearity of expectation and Lemma 5.

Clearly,

\[
E \left[ \left( \sum_{i=1}^{|B(z, \delta_{nr})|/2} d(\pi(2i - 1), \pi(2i)) \right)^2 \right]
\]

\[
= E \left[ \sum_{i=1}^{|B(z, \delta_{nr})|/2} d(\pi(2i - 1), \pi(2i)) \cdot \sum_{j=1}^{|B(z, \delta_{nr})|/2} d(\pi(2j - 1), \pi(2j)) \right]
\]

\[
= \sum_{\text{distinct } i, j = 1} \sum_{|B(z, \delta_{nr})|/2} E \left[ d(\pi(2i - 1), \pi(2i)) \cdot d(\pi(2j - 1), \pi(2j)) \right] + \sum_{i=1} \sum_{|B(z, \delta_{nr})|/2} E \left[ d^2(\pi(2i - 1), \pi(2i)) \right],
\]

where the last equality follows from the linearity of expectation and the separation of pairs \((i, j)\) according to whether \(i = j\).

Lemma 8. With expectations taken over \( \pi \),

\[
\sum_{\text{distinct } i, j = 1} \sum_{|B(z, \delta_{nr})|/2} E \left[ d(\pi(2i - 1), \pi(2i)) \cdot d(\pi(2j - 1), \pi(2j)) \right] \leq \frac{1}{4} \cdot (1 + o(1)) n^2 (r')^2.
\]

Proof. Pick any distinct \( i, j \in \lfloor |B(z, \delta_{nr})|/2 \rfloor \). By line 5 of Las Vegas median, \( \{\pi(2i - 1), \pi(2i), \pi(2j - 1), \pi(2j)\} \) is a uniformly random size-4 subset of \( B(z, \delta_{nr}) \). So

\[
E \left[ d(\pi(2i - 1), \pi(2i)) \cdot d(\pi(2j - 1), \pi(2j)) \right] = \frac{1}{|B(z, \delta_{nr})| \cdot (|B(z, \delta_{nr})| - 1) \cdot (|B(z, \delta_{nr})| - 2) \cdot (|B(z, \delta_{nr})| - 3)} \sum_{\text{distinct } u, v, x, y \in B(z, \delta_{nr})} d(u, v) \cdot d(x, y).
\]
Clearly,
\[
\sum_{\text{distinct } u, v, x, y \in B(z, \delta nr)} d(u, v) \cdot d(x, y) \leq \sum_{u,v,x,y \in B(z, \delta nr)} d(u, v) \cdot \sum_{x,y \in B(z, \delta nr)} d(x, y)
\]
\[
= \left( \sum_{u,v \in B(z, \delta nr)} d(u, v) \right)^2.
\]

In summary,
\[
\lceil |B(z, \delta nr)|/2 \rceil \sum_{i, j = 1; \text{distinct}} E \left[ d^2(2i - 1, \pi(2i)) \cdot d(2j - 1, \pi(2j)) \right] \leq \left( \frac{|B(z, \delta nr)|}{2} \right) \left( \left( \frac{|B(z, \delta nr)|}{2} \right) - 1 \right)
\]
\[
\cdot \frac{|B(z, \delta nr)| \cdot (|B(z, \delta nr)| - 1) \cdot (|B(z, \delta nr)| - 2) \cdot (|B(z, \delta nr)| - 3)}{\sum_{u,v \in B(z, \delta nr)} d(u, v)}.
\]

Together with Lemma 5 and equation (9), this completes the proof. \qed

**Lemma 9.** With expectations taken over \( \pi \),
\[
\lceil |B(z, \delta nr)|/2 \rceil \sum_{i=1}^{\lceil |B(z, \delta nr)|/2 \rceil} E \left[ d^2(2i - 1, \pi(2i)) \right] \leq (1 + o(1)) \left( \delta n^2 r r' + 2 \delta^2 n r^2 \right). \tag{17}
\]

**Proof.** By line 5 of Las Vegas median, \( \{ \pi(2i - 1), \pi(2i) \} \) is a uniformly random size-2 subset of \( B(z, \delta nr) \) for each \( i \in \lceil |B(z, \delta nr)|/2 \rceil \). Therefore,
\[
\lceil |B(z, \delta nr)|/2 \rceil \sum_{i=1}^{\lceil |B(z, \delta nr)|/2 \rceil} E \left[ d^2(2i - 1, \pi(2i)) \right] \leq (1 + o(1)) \left( \delta n^2 r r' + 2 \delta^2 n r^2 \right). \tag{18}
\]
For all \( u, v \in B(z, \delta nr) \),

\[
d(u, v) \leq d(z, u) + d(z, v) \leq \delta nr + \delta nr = 2\delta nr,
\]

(19)

where the first inequality follows from the triangle inequality.

By equations (9) and (18)–(19), the left-hand side of inequality (17) cannot exceed the optimal value of the following problem, called MAX SQUARE SUM:

Find \( d_{u,v} \in \mathbb{R} \) for all \( u, v \in B(z, \delta nr) \) to maximize

\[
\frac{\lfloor |B(z, \delta nr)| \rfloor}{2} \cdot \frac{1}{|B(z, \delta nr)| \cdot (|B(z, \delta nr)| - 1)} \cdot \sum_{u,v \in B(z, \delta nr)} d_{u,v}^2
\]

(20)

subject to

\[
\frac{1}{|B(z, \delta nr)|^2} \cdot \sum_{u,v \in B(z, \delta nr)} d_{u,v} = r',
\]

(21)

\forall u, v \in B(z, \delta nr), 0 \leq d_{u,v} \leq 2\delta nr.

(22)

Above, constraint (21) (resp., (22)) mimics equation (9) (resp., inequality (19) and the non-negativeness of distances). Appendix A bounds the optimal value of MAX SQUARE SUM from above by

\[
\frac{|B(z, \delta nr)|}{2} \cdot \frac{1}{|B(z, \delta nr)| \cdot (|B(z, \delta nr)| - 1)} \cdot \left( \left\lfloor \frac{|B(z, \delta nr)|^2 r'}{2\delta nr} \right\rfloor + 1 \right) \cdot (2\delta nr)^2.
\]

This evaluates to be at most \((1 + o(1))(\delta n^2 r r' + 2\delta^2 n r^2)\) by Lemma 5.

Recall that the variance of any random variable \( X \) equals \( \text{E}[X^2] - (\text{E}[X])^2 \).

**Lemma 10.** With variances taken over \( \pi \),

\[
\text{var} \left( \sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i-1), \pi(2i)) \right) \leq 2 (1 + o(1)) \delta n^2 r^2.
\]

**Proof.** By equations (15)–(16) and Lemmas 8–9,

\[
\text{E} \left( \left( \sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i-1), \pi(2i)) \right)^2 \right) \leq \frac{1}{4} \cdot (1 + o(1)) n^2 (r')^2 + (1 + o(1)) (\delta^2 r r' + 2\delta^2 n r^2).
\]

This and Lemma 7 imply

\[
\text{var} \left( \sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i-1), \pi(2i)) \right) \leq o(1) \cdot n^2 (r')^2 + (1 + o(1)) (\delta^2 r r' + 2\delta^2 n r^2).
\]

Finally, invoke Lemma 6. \( \square \)
Lemma 11. For all \( k > 1 \),
\[
\Pr \left[ \left| \sum_{i=1}^{[|B(z,\delta nr)|/2]} d(\pi(2i-1), \pi(2i)) - \frac{1}{2} \cdot (1 \pm o(1)) nr' \right| \geq k \sqrt{2 (1 + o(1)) \delta nr} \right] \leq \frac{1}{k^2},
\]
where the probability is taken over \( \pi \).

Proof. Use Chebyshev’s inequality and Lemmas 7 and 10. \( \square \)

Let \( z' \in B(z, \delta nr) \) be a 1-median of \( B(z, \delta nr) \), i.e.,
\[
z' = \arg\min_{y \in B(z, \delta nr)} \sum_{x \in B(z, \delta nr)} d(y, x),
\]
breaking ties arbitrarily. So by the averaging argument,
\[
\sum_{x \in B(z, \delta nr)} d(z', x) \leq \frac{1}{|B(z, \delta nr)|} \cdot \sum_{y \in B(z, \delta nr)} \sum_{x \in B(z, \delta nr)} d(y, x). \tag{23}
\]

Lemma 12.
\[
\sum_{x \in B(z, \delta nr)} d(z', x) \leq nr'.
\]

Proof. We have
\[
\sum_{x \in B(z, \delta nr)} d(z', x) \leq \frac{1}{|B(z, \delta nr)|} \cdot \sum_{u,v \in B(z, \delta nr)} d(u,v) \quad = \quad |B(z, \delta nr)| \cdot r'. \tag{9}
\]
Clearly, \(|B(z, \delta nr)| \leq n\). \( \square \)

Lemma 13. For all sufficiently large \( n \),
\[
d(z', z) \leq 8r.
\]

Proof. We have
\[
\sum_{x \in B(z, \delta nr)} d(z', x) \geq \sum_{x \in B(z, \delta nr)} (d(z', z) - d(z, x)) \tag{24}
\geq \sum_{x \in B(z, \delta nr)} d(z', z) - \sum_{x \in [n]} d(z, x)
= \left( \sum_{x \in B(z, \delta nr)} d(z', z) \right) - nr
= \left| B(z, \delta nr) \right| \cdot d(z', z) - nr,
\]
where the first inequality (resp., the first equality) follows from the triangle inequality (resp., line 4 of Las Vegas median). By Lemmas 6 and 12,
\[ \sum_{x \in B(z, \delta nr)} d(z', x) \leq 2nr. \] (25)

By inequalities (24)–(25) and Lemma 5, \( d(z', z) \leq (3 + o(1))r. \) \( \square \)

**Lemma 14.** For all sufficiently large \( n \),
\[ \sum_{x \in [n]} d(z', x) \leq nr' + \frac{16r}{\delta} + \sum_{x \in [n] \setminus B(z, \delta nr)} (d(z, x) - 8r). \]

**Proof.** By the triangle inequality,
\[ \sum_{x \in [n] \setminus B(z, \delta nr)} d(z', x) \leq \sum_{x \in [n] \setminus B(z, \delta nr)} (d(z', z) + d(z, x)) \leq \sum_{x \in [n] \setminus B(z, \delta nr)} (8r + d(z, x)) \leq \frac{16r}{\delta} + \sum_{x \in [n] \setminus B(z, \delta nr)} (d(z, x) - 8r). \]

Now sum up the above with the inequality in Lemma 12. \( \square \)

**Lemma 15.** For all sufficiently large \( n \) and with probability greater than \( 1/2 \),
\[ \sum_{i=1}^{|[B(z, \delta nr)]/2} d(\pi(2i - 1), \pi(2i)) + \sum_{x \in [n] \setminus B(z, \delta nr)} (d(z, x) - 8r) \geq \frac{1 - 100\sqrt{\delta}}{2} \cdot nr. \] (26)

where the probability is taken over \( \pi \) and the internal coin tosses of Indyk median in line 3 of Las Vegas median.

**Proof.** By Lemma 11 with \( k = 5 \),
\[ \sum_{i=1}^{|[B(z, \delta nr)]/2} d(\pi(2i - 1), \pi(2i)) > \frac{1}{2} \cdot (1 + o(1))nr' - \frac{5}{2} \sqrt{2} (1 + o(1)) \delta nr. \] (27)

with probability at least \( 1 - 1/25 \). By Fact 1 and line 3 of Las Vegas median,
\[ \sum_{x \in [n]} d(z, x) \leq \left(1 + \frac{\epsilon}{10^{10}}\right) \cdot \min_{y \in [n]} \sum_{x \in [n]} d(y, x) \leq \left(1 + \frac{\epsilon}{10^{10}}\right) \cdot \sum_{x \in [n]} d(z', x) \] (28) (29)

\(^5\)In fact, this is stronger than the lemma to be proved.
with probability at least $1 - 1/e$. Now by the union bound, inequalities (27)–(29) hold simultaneously with probability at least $1 - 1/25 - 1/e > 1/2$. It remains to derive inequality (26) from inequalities (27)–(29) for all sufficiently large $n$.

Line 4 of Las Vegas median, inequalities (28)–(29) and Lemma 14 give

$$nr \leq \left(1 + \frac{\epsilon}{10^{10}}\right) \left(2 (1 \pm o(1)) \left(\sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i-1), \pi(2i))\right) + 5\sqrt{2 (1 + o(1)) \delta nr}\right) + \frac{16r}{\delta} + \sum_{x \in [n] \setminus B(z, \delta nr)} (d(z, x) - 8r).$$

This and inequality (27) imply

$$nr \leq \left(1 + \frac{\epsilon}{10^{10}}\right) \left(2 (1 \pm o(1)) \left(\sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i-1), \pi(2i))\right) + 5\sqrt{2 (1 + o(1)) \delta nr}\right) + \frac{16r}{\delta} + \sum_{x \in [n] \setminus B(z, \delta nr)} (d(z, x) - 8r).$$

6 Clearly, $16r/\delta \leq 0.01 \cdot \sqrt{\delta} nr$ for all sufficiently large $n$. So inequality (31) implies, for all sufficiently large $n$ and after laborious calculations,

$$nr - \left(1 + \frac{\epsilon}{10^{10}}\right) \left(2 (1 \pm o(1)) \left(\sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i-1), \pi(2i))\right) + 5\sqrt{2 (1 + o(1)) \delta nr}\right) + \frac{16r}{\delta} + \sum_{x \in [n] \setminus B(z, \delta nr)} (d(z, x) - 8r).$$

This implies inequality (26) for all sufficiently large $n$ (note that $\epsilon/10^{10} < \sqrt{\delta}$ by line 1 of Las Vegas Median). 7

Lemma 15 and lines 6–7 of Las Vegas median show the probability of termination in any iteration to be $\Omega(1)$. Because the proof of Lemma 15 implies that inequalities (26)–(29) hold simultaneously with probability $\Omega(1)$ in any iteration of Las Vegas median, it happens with probability $\Omega(1)$ that in the first iteration, $z$ is returned in line 7 (because of inequality (26)) and is $(1 + \epsilon/10^{10})$-approximate (because of inequality (28)). So Las Vegas median outputs a $(1 + \epsilon/10^{10})$-approximate 1-median with probability $\Omega(1)$ in the first iteration. In summary, we have the following.

---

6 To see this, rewrite inequality (27) as

$$nr' < 2 (1 \pm o(1)) \left(\sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i-1), \pi(2i))\right) + 5\sqrt{2 (1 + o(1)) \delta nr}\right).$$

7 Divide both sides by $(2 + 2\epsilon/10^{10})(1 + o(1))$ so that the coefficient before $\sum_{i=1}^{\lfloor |B(z, \delta nr)|/2 \rfloor} d(\pi(2i-1), \pi(2i))$ becomes 1 in the right-hand side. Then verify the left-hand side (which is now $(nr - (1 + \epsilon/10^{10})\sqrt{2 (1 + o(1)) \delta nr})(2 + 2\epsilon/10^{10})(1 + o(1))$) to be at least $(1 - 100/\sqrt{\delta})nr/2$ for all sufficiently large $n$.
Lemma 16. The first iteration of the \textbf{while} loop of Las Vegas median outputs a \((1 + \epsilon)\)-approximate 1-median with probability \(\Omega(1)\).

5 Putting things together

We now show that \textsc{Metric 1-Median} has a Las Vegas \((2 + \epsilon)\)-approximation algorithm with an expected \(O(n/\epsilon^2)\) running time for all constants \(\epsilon > 0\). Our algorithm also outputs a \((1 + \epsilon)\)-approximate 1-median in time \(O(n/\epsilon^2)\) with probability \(\Omega(1)\).

Theorem 17. For each constant \(\epsilon > 0\), \textsc{Metric 1-Median} has a randomized algorithm that (1) always outputs a \((2 + \epsilon)\)-approximate solution in an expected \(O(n/\epsilon^2)\) time and that (2) outputs a \((1 + \epsilon)\)-approximate solution in time \(O(n/\epsilon^2)\) with probability \(\Omega(1)\).

Proof. By Lemma 4, Las Vegas median outputs a \((2+\epsilon)\)-approximate 1-median at termination. To prevent Las Vegas median from running forever, find a 1-median by brute force (which obviously takes \(O(n^2)\) time) after \(n^2\) steps of computation.

By Fact 1, line 3 of Las Vegas median takes \(O(n/\epsilon^2)\) time. Line 5 takes time \(O(|B(z, \delta nr)|) = O(n)\) by the Knuth shuffle. Clearly, the other lines also take \(O(n)\) time. Consequently, each iteration of the \textbf{while} loop of Las Vegas median takes \(O(n/\epsilon^2)\) time. By Lemma 15 and lines 6–7, Las Vegas median runs for at most \(1/\Omega(1) = O(1)\) iterations in expectation. So its expected running time is \(O(1) \cdot O(n/\epsilon^2) = O(n/\epsilon^2)\).

Having shown each iteration of Las Vegas median to take \(O(n/\epsilon^2)\) time, establish condition (2) of the theorem with Lemma 16.

By Fact 1, Indyk median satisfies condition (2) in Theorem 17. But it does not satisfy condition (1).

We briefly justify the optimality of the ratio of \(2 + \epsilon\) in Theorem 17. Let \(A\) be a randomized algorithm that always outputs a \((2 - \epsilon)\)-approximate 1-median. Furthermore, denote by \(p \in [n]\) (resp., \(Q \subseteq [n] \times [n]\)) the output (resp., the set of queries as unordered pairs) of \(A^{d_1}(n)\), where \(d_1\) is the discrete metric (i.e., \(d_1(x, y) = 1\) and \(d_1(x, x) = 0\) for all distinct \(x, y \in [n]\)). Without loss of generality, assume \((p, y) \in Q\) for all \(y \in [n] \setminus \{p\}\) by adding dummy queries. So \(A\) knows that

\[
\sum_{y \in [n]} d_1(p, y) = n - 1. \tag{32}
\]

Furthermore, assume that \(A\) never queries for the distance from a point to itself.

In the sequel, consider the case that \(|Q| < \epsilon \cdot (n - 1)^2/4\). By the averaging argument, there exists a point \(\hat{p} \in [n] \setminus \{p\}\) involved in at most \(2 \cdot |Q|/(n - 1)\)
queries in $Q$. Clearly, $A$ cannot exclude the possibility that $d_1(\hat{p}, y) = 1/2$ for all $y \in [n] \setminus \{\hat{p}\}$ satisfying $(\hat{p}, y) \notin Q$. In summary, $A$ cannot rule out the case that

$$\sum_{y \in [n]} d_1(\hat{p}, y) \leq \frac{2 \cdot |Q|}{n - 1} \cdot 1 + \left( n - 1 - \frac{2 \cdot |Q|}{n - 1} \right) \cdot \frac{1}{2} < \left( \frac{1}{2} + \frac{\epsilon}{4} \right) \cdot (n - 1). \quad (33)$$

Equations (32)–(33) contradict the guarantee that $p$ is $(2 - \epsilon)$-approximate. In summary, any randomized algorithm that always outputs a $(2 - \epsilon)$-approximate 1-median must always make at least $\epsilon \cdot (n - 1)^2 / 4 = \Omega(\epsilon n^2)$ queries given oracle access to the discrete metric.

A Analyzing MAX SQUARE SUM

MAX SQUARE SUM has an optimal solution, denoted $\{\tilde{d}_{u,v} \in \mathbb{R}\}_{u,v \in B(z, \delta nr)}$, because its feasible solutions (i.e., those satisfying constraints (21)–(22)) form a closed and bounded subset of $\mathbb{R}^{|B(z, \delta nr)^2|}$. (Recall from elementary mathematical analysis that a continuous real-valued function on a closed and bounded subset of $\mathbb{R}^k$ has a maximum value, where $k < \infty$.) Note that $\{\tilde{d}_{u,v} \in \mathbb{R}\}_{u,v \in B(z, \delta nr)}$ must be feasible to MAX SQUARE SUM. Below is a consequence of constraint (21).

Lemma A.1.

$$\left| \left\{ (u, v) \in B^2(z, \delta nr) \mid \tilde{d}_{u,v} = 2\delta nr \right\} \right| \leq \left\lfloor \frac{|B(z, \delta nr)|^2 r'}{2\delta nr} \right\rfloor. \quad (34)$$

Proof. Clearly,

$$|B(z, \delta nr)|^2 r' \overset{(21)}{=} \sum_{u,v \in B(z, \delta nr)} \tilde{d}_{u,v} \geq \left| \left\{ (u, v) \in B^2(z, \delta nr) \mid \tilde{d}_{u,v} = 2\delta nr \right\} \right| \cdot 2\delta nr.$$

Furthermore, the left-hand side of inequality (34) is an integer. \qed

Lemma A.2.

$$\left| \left\{ (u, v) \in B^2(z, \delta nr) \mid \tilde{d}_{u,v} > 0 \right\} \right| \leq \left\lfloor \frac{|B(z, \delta nr)|^2 r'}{2\delta nr} \right\rfloor + 1.$$

Proof. Assume otherwise. Then

$$\left| \left\{ (u, v) \in B^2(z, \delta nr) \mid (\tilde{d}_{u,v} > 0) \land (\tilde{d}_{u,v} \neq 2\delta nr) \right\} \right| \geq \left| \left\{ (u, v) \in B^2(z, \delta nr) \mid \tilde{d}_{u,v} > 0 \right\} \right| - \left| \left\{ (u, v) \in B^2(z, \delta nr) \mid \tilde{d}_{u,v} = 2\delta nr \right\} \right| \geq \frac{|B(z, \delta nr)|^2 r'}{2\delta nr} + 2 - \left| \left\{ (u, v) \in B^2(z, \delta nr) \mid \tilde{d}_{u,v} = 2\delta nr \right\} \right| \overset{\text{Lemma A.1}}{\geq} 2.$$
So by constraint (22) (and the feasibility of \( \{ \tilde{d}_{u,v} \}_{u,v \in B(z, \delta nr)} \) to MAX SQUARE SUM),
\[
\left| \left\{ (u, v) \in B^2(z, \delta nr) \mid 0 < \tilde{d}_{u,v} < 2\delta nr \right\} \right| \geq 2.
\]
Consequently, there exist distinct \((x, y), (x', y') \in B^2(z, \delta nr)\) satisfying
\[
0 < \tilde{d}_{x,y}, \tilde{d}_{x',y'} < 2\delta nr.
\]
(35)

By symmetry, assume \(\tilde{d}_{x,y} \geq \tilde{d}_{x',y'}\). By inequality (35), there exists a small real number \(\beta > 0\) such that increasing \(\tilde{d}_{x,y}\) by \(\beta\) and simultaneously decreasing \(\tilde{d}_{x',y'}\) by \(\beta\) will preserve constraints (21)–(22). I.e., the solution \(\{ \hat{d}_{u,v} \in \mathbb{R} \}_{u,v \in B(z, \delta nr)}\) defined below is feasible to MAX SQUARE SUM:
\[
\hat{d}_{u,v} = \begin{cases} 
\tilde{d}_{x,y} + \beta, & \text{if } (u, v) = (x, y), \\
\tilde{d}_{x',y'} - \beta, & \text{if } (u, v) = (x', y'), \\
\tilde{d}_{u,v}, & \text{otherwise}.
\end{cases}
\]
(36)

Clearly, objective (20) w.r.t. \(\{ \hat{d}_{u,v} \}_{u,v \in B(z, \delta nr)}\) exceeds that w.r.t. \(\{ \tilde{d}_{u,v} \}_{u,v \in B(z, \delta nr)}\) by
\[
\left\lfloor \frac{|B(z, \delta nr)|}{2} \right\rfloor \cdot \frac{1}{|B(z, \delta nr)| \cdot (|B(z, \delta nr)| - 1)} \cdot \sum_{u,v \in B(z, \delta nr)} \left( \hat{d}_{u,v}^2 - \tilde{d}_{u,v}^2 \right)
\]
(36)

\[
\geq \left[ \frac{|B(z, \delta nr)|}{2} \right] \cdot \frac{1}{|B(z, \delta nr)| \cdot (|B(z, \delta nr)| - 1)} \cdot \left( (\tilde{d}_{x,y} + \beta)^2 + (\tilde{d}_{x',y'} - \beta)^2 - \tilde{d}_{x,y}^2 - \tilde{d}_{x',y'}^2 \right)
\]
\[
= \left[ \frac{|B(z, \delta nr)|}{2} \right] \cdot \frac{1}{|B(z, \delta nr)| \cdot (|B(z, \delta nr)| - 1)} \cdot \left( 2\beta \tilde{d}_{x,y} - 2\beta \tilde{d}_{x',y'} + 2\beta^2 \right)
\]
\[
> 0,
\]

where the inequality holds because \(\tilde{d}_{x,y} \geq \tilde{d}_{x',y'}\) and \(\beta > 0\).

In summary, \(\{ \hat{d}_{u,v} \}_{u,v \in B(z, \delta nr)}\) is a feasible solution achieving a greater objective (20) than the optimal solution \(\{ \tilde{d}_{u,v} \}_{u,v \in B(z, \delta nr)}\) does, a contradiction. \(\Box\)

We now bound the optimal value of MAX SQUARE SUM.

**Theorem A.3.** The optimal value of MAX SQUARE SUM is at most
\[
\left[ \frac{|B(z, \delta nr)|}{2} \right] \cdot \frac{1}{|B(z, \delta nr)| \cdot (|B(z, \delta nr)| - 1)} \cdot \left( \left[ \frac{|B(z, \delta nr)|^2 r'}{2\delta nr} \right] + 1 \right) \cdot (2\delta nr)^2
\]
Proof. W.r.t. the optimal (and thus feasible) solution $\{\tilde{d}_{u,v}\}_{u,v\in B(z,\epsilon nr)}$, objective (20) equals
\[
\frac{|B(z,\delta nr)|}{2} \cdot \frac{1}{|B(z,\delta nr)| \cdot (|B(z,\delta nr)| - 1)} \cdot \sum_{u,v\in B(z,\delta nr)} \chi[\tilde{d}_{u,v} \neq 0] \cdot \tilde{d}_{u,v}^2
\]
\[
\leq \frac{|B(z,\delta nr)|}{2} \cdot \frac{1}{|B(z,\delta nr)| \cdot (|B(z,\delta nr)| - 1)} \cdot \sum_{u,v\in B(z,\delta nr)} \chi[\tilde{d}_{u,v} > 0] \cdot (2\delta nr)^2,
\]
where $\chi[P] = 1$ if $P$ is true and $\chi[P] = 0$ otherwise, for any predicate $P$. Now invoke Lemma A.2.

\[\square\]

References

[1] C.-L. Chang. Deterministic sublinear-time approximations for metric 1-median selection. Information Processing Letters, 113(8):288–292, 2013.

[2] C.-L. Chang. A deterministic sublinear-time nonadaptive algorithm for metric 1-median selection. Theoretical Computer Science, 602:149–157, 2015.

[3] C.-L. Chang. Metric 1-median selection: Query complexity vs. approximation ratio. In Proceedings of the 22nd International Computing and Combinatorics Conference, pages 131–142, Ho Chi Minh City, Vietnam, 2016. Full version at https://arxiv.org/abs/1509.05662.

[4] C.-L. Chang. A lower bound for metric 1-median selection. Journal of Computer and System Sciences, 84:44–51, 2017.

[5] D. Eppstein and J. Wang. Fast approximation of centrality. Journal of Graph Algorithms and Applications, 8(1):39–45, 2004.

[6] O. Goldreich and D. Ron. Approximating average parameters of graphs. Random Structures & Algorithms, 32(4):473–493, 2008.

[7] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O’Callaghan. Clustering data streams: Theory and practice. IEEE Transactions on Knowledge and Data Engineering, 15(3):515–528, 2003.

[8] P. Indyk. Sublinear time algorithms for metric space problems. In Proceedings of the 31st Annual ACM Symposium on Theory of Computing, pages 428–434, 1999.

[9] P. Indyk. High-dimensional computational geometry. PhD thesis, Stanford University, 2000.
[10] A. Kumar, Y. Sabharwal, and S. Sen. Linear-time approximation schemes for clustering problems in any dimensions. *Journal of the ACM*, 57(2):5, 2010.

[11] R. R. Mettu and C. G. Plaxton. Optimal time bounds for approximate clustering. *Machine Learning*, 56(1–3):35–60, 2004.

[12] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, UK, 1995.

[13] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition, 1976.

[14] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[15] B.-Y. Wu. On approximating metric 1-median in sublinear time. *Information Processing Letters*, 114(4):163–166, 2014.