Better Bounds for Incremental Medians

Marek Chrobak[*] Mathilde Hurand[†]

Abstract

In the incremental version of the well-known *k-median problem* the objective is to compute an incremental sequence of facility sets $F_1 \subseteq F_2 \subseteq .... \subseteq F_n$, where each F_k contains at most k facilities. We say that this incremental medians sequence is *R-competitive* if the cost of each F_k is at most R times the optimum cost of k facilities. The smallest such R is called the *competitive ratio* of the sequence $\{F_k\}$. Mettu and Plaxton [6, 7] presented a polynomial-time algorithm that computes an incremental sequence with competitive ratio ≈ 30. They also showed a lower bound of 2. The upper bound on the ratio was improved to 8 in [5] and [4]. We improve both bounds in this paper. We first show that no incremental sequence can have competitive ratio better than 2.01 and we give a probabilistic construction of a sequence whose competitive ratio is at most $2 + 4\sqrt{2} \approx 7.656$. We also propose a new approach to the problem that for instances that we refer to as *equable* achieves an optimal ratio of 2.

1 Introduction

The *k-median* problem is one of the most studied facility location problems. We are given two sets: a set $\mathcal{C}$ of *customers* and a set $\mathcal{F}$ of n *facilities*, with a metric function d that specifies the distance d_{xy} between any two points $x, y \in \mathcal{C} \cup \mathcal{F}$. The cost of a facility set $F \subseteq \mathcal{F}$, denoted by $cost(F)$, is defined as the minimum sum, over all customers $c \in \mathcal{C}$, of d_{cF}, where $d_{cF} = \min_{f \in F} d_{cf}$ is the minimum distance from c to F. Given k, the objective is to compute a set of k facilities with minimum cost.

Not surprisingly, the k-median problem is NP-hard. A number of polynomial-time approximation algorithms have been proposed, with the latest one, by Arya *et al.* [1, 2] achieving the ratio of $3 + \epsilon$, for any $\epsilon > 0$.

Mettu and Plaxton [6, 7] introduced the *incremental medians problem*, where the permitted number k of facilities is not specified in advance. Starting with the empty set, an algorithm receives authorizations for new facilities over time, and after each authorization it is allowed to add another facility to the existing ones. As a result, such an algorithm produces an incremental sequence of facility sets $F_1 \subseteq F_2 \subseteq ... \subseteq F_n$, where $|F_k| \leq k$ for all k. This sequence $\{F_k\}$ is said to be *R-competitive* if $cost(F_k)$ is at most R times the optimum cost of k facilities, for each k. The smallest such R is called the *competitive ratio* of $\{F_k\}$.

Mettu and Plaxton [6, 7] gave a polynomial-time algorithm that computes such an incremental sequence with competitive ratio ≈ 30. This result is quite remarkable, for there is no apparent

[*] Department of Computer Science, University of California, Riverside, CA 92521. Research supported by NSF Grants CCR-0208856, OISE-0340752, and CCF-0729071.

[†] Department d'Informatique (LIX), Ecole Polytechnique, Palaiseau, France.

reason why an incremental sequence $\{F_k\}$ of facility sets, with each $\mathit{cost}(F_k)$ within a constant factor of the the optimum, would even exist – let alone be computed efficiently.

It is thus natural to address the issue of *existence* separately from *computational complexity*, and this is what we focus on in this paper. As shown by Mettu and Plaxton [6, 7], no ratio better than 2 is possible, that is, for each $\epsilon > 0$ there is a metric space where each incremental facility sequence has competitive ratio at least $2 - \epsilon$. The upper bound on the ratio was improved to 8 by Lin *et al.* [5] and, independently, by Chrobak *et al.* [4]. In [5], the authors also show that a 16-competitive incremental median sequence can be computed in polynomial time.

Our results. We improve both the lower and upper bounds for incremental medians. For the lower bound, we show that, in general, no competitive ratio better than 2.01 is possible. We also prove, via a probabilistic argument, that each instance has an incremental medians sequence with competitive ratio at most $2 + 4\sqrt{2} \approx 7.656$.

In numerical terms, the improvement of the lower bound is mostly symbolic, as it implies that 2 is not the "right" ratio. For the upper bound, our result shows that the doubling method from [5, 4] (see also [3]) is not optimal – even though it gives the optimal ratio of 4 for the closely related "resource augmentation" version of incremental medians [4]. As discussed in Section 6, we believe that our methods can be refined to further improve both the lower and upper bounds.

In addition, we consider a special case of the incremental medians problem where for any fixed value of k, each customer has the same distance to the optimal k-median. We refer to such instances as *equable*. (See Section 5 for a formal definition.) For this case, we show a construction of a 2-competitive incremental medians sequence, matching the lower bound from [6, 7]. Our method for this case is very different from previous constructions and we believe that it will be useful in improving the upper bound for general spaces. In fact, this result implies that if there is a constant $\gamma \geq 1$ such that for each fixed k all customers' optimal costs are within factor γ of each other, then our construction achieves ratio at most 2γ – improving our own bound above if $\gamma < 1 + 2\sqrt{2}$.

2 Preliminaries

Let $(\mathcal{F}, \mathcal{C})$ be an instance of the medians problem, where $\mathcal{F}$ is a set of n facilities, $\mathcal{C}$ is a set of customers, and $\mathcal{F} \cup \mathcal{C}$ forms a metric space. By d_{xy} or $d(x, y)$ we denote the distance between points x, y. If Y is a set, we also write $d_{xY} = \min_{y \in Y} d_{xy}$ for the minimum distance from x to Y. For a facility set $F \subseteq \mathcal{F}$, denote by $\mathit{cost}(F)$ the cost of F, that is $\sum_{x \in \mathcal{C}} d_{xF}$. We will simplify the notation for cost when F has small cardinality by omitting set notation and writing $\mathit{cost}(x) = \mathit{cost}(\{x\})$, $\mathit{cost}(x, y) = \mathit{cost}(\{x, y\})$, etc., for $x, y \in \mathcal{F}$.

For a point x and a set Y, denote by $\Gamma_Y(x)$ the point $y \in Y$ that is closest to x, that is $d_{xy} = d_{xY}$ (if this point is not unique, then break the tie arbitrarily.) If X is a set, we also define $\Gamma_Y(X) = \{\Gamma_Y(x) \mid x \in X\}$. Clearly, $|\Gamma_Y(X)| \leq |X|$. Note that if F is a facility set and X is a set of customers, then $\Gamma_F(X)$ is exactly the set of facilities in F that serve customers in X if F is the facility set under consideration.

By opt_k we denote the optimum cost of k facilities, that is

$$\mathit{opt}_k \;=\; \min\{\mathit{cost}(F) \mid F \subseteq \mathcal{F} \text{ and } |F| = k\}. \tag{1}$$

By $F^*_k \subseteq \mathcal{F}$ we will denote the optimal set of k facilities, that is, the *k-median.* (As before, ties are broken arbitrarily.) Thus $\mathit{cost}(F^*_k) = \mathit{opt}_k$.

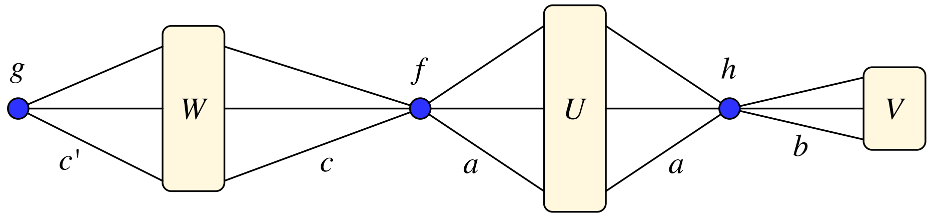

Figure 1: Metric space used in the lower bound. The varying lengths and rectangle sizes represent, approximately, relative distances and set cardinalities.

3 A New Lower Bound

In this section we prove our lower bound of 2.01 on the competitive ratio for incremental medians, improving slightly the previous bound of 2 from [6, 7].

Theorem 1. *There is an instance $(\mathcal{C}, \mathcal{F})$ for which no incremental median sequence has competitive ratio smaller than* 2.01*.*

Proof. In our construction, the set of customers is $\mathcal{C} = U \cup V \cup W$, where U, V, W are disjoint sets with $|U| + |V| + |W| = n - 3$, for some large integer n. The set of facilities is $\mathcal{F} = \{f, g, h\} \cup \mathcal{C}$. The distances between customers and facilities are illustrated in Figure 1. A bi-directional edge between a facility f, g or h and a set U, V or W signifies that this facility is connected to all customers in this set by an edge of the indicated distance. Thus, for each set U, V, W, all customers in a set have the same distance to each facility. For example, the distance from f to all $u \in U$ is a, the distance from h to all $v \in V$ is b, etc. Other distances are measured along the shortest paths in the graph represented in Figure 1. For instance, the distance from g to h is $c' + c + 2a$, the distance from f to any $v \in V$ is $2a + b$. The same rule applies, in particular, to any two customers from a same set (they are *not* at distance 0 from one-another). For example, for $v, v' \in V$ with $v' \neq v$, the distance from v to v' is $2b$, for $w, w' \in W$ with $w' \neq w$, the distance from w to w' is $2 \min\{c, c'\}$.

Since for $k = n - 3$ the optimal cost is 0, the first $n - 3$ facilities in any competitive incremental sequence must be chosen from $\mathcal{C}$. In fact, we will only use only three values of k: $k = 1, 2$ and $n - 3$.

To prove that there is no incremental medians sequence with ratio better than R, we only need to give some values a, b, c, c', $|U|$, $|V|$ and $|W|$ such that:

$$\min\{\mathit{cost}(v), \mathit{cost}(w)\} \geq R \cdot \mathit{cost}(f) \quad \text{and} \tag{2}$$

$$\min\{\mathit{cost}(u, u'), \mathit{cost}(u, v), \mathit{cost}(u, w)\} \geq R \cdot \mathit{cost}(g, h) \tag{3}$$

for any $u, u' \in U$, $v \in V$ and $w \in W$.

These inequalities imply the lower bound of R, because (2) implies that, for $k = 1$, to beat ratio R we must pick some $u \in U$ as the first facility, and (3) implies that, for $k = 2$, it is not possible to add to u another facility and preserve ratio R.

In order to simplify calculations, we slightly modify the way we compute the costs. If $x \in U \cup V \cup W$ is chosen as a facility and it serves a customer at a point $z \neq x$ from the same set U, V or W, then the cost of z is the length of the shortest path from z to x via one facility f, g, or h, while the cost of $z = x$ is 0. Our modification is that we will charge this $z = x$ the cost of such a shortest path as well, that is, z cannot serve itself directly at cost 0. Thus, if there is a facility at $x \in W$, then we will charge x the cost of $2\min\{c, c'\}$ to get to this facility; if $x \in U$, this cost will be $2a$, and if $x \in V$ this cost will be $2b$. Let $cost'(\cdot)$ denote this modified cost function. Note that for any facility set F of constant cardinality, we have $cost'(F) = (1 + \Theta(1/n))cost(F)$, since all customers, except those located at the points of F, contribute the same cost to both cost functions; thus for $k = 1, 2$ and n large enough, the two cost functions are essentially identical. Further, if $F \subseteq \{f, g, h\}$, then $cost'(F) = cost(F)$.

With this convention in mind, we set $a = 5/4$, $b = 1$, $c = 211/100$, $c' = 141/100$, $|U| = 295\lambda$, $|V| = 25\lambda$, and $|W| = 149\lambda$, for some large integer λ. (Thus $n = 469\lambda{+}3$.) Note that $b \le a \le c' \le c$.

Fix any $u, u' \in U$, with $u \ne u'$, $v \in V$ and $w \in W$. Then, for $k = 1$ we have

$$\begin{aligned} cost(f) &= |U|a + |V|(b + 2a) + |W|c \;=\; 776.64\lambda \\ cost'(v) &= |U|(a + b) + |V|(2b) + |W|(b + 2a + c) \;=\; 1549.64\lambda \\ cost'(w) &= |U|(a + c) + |V|(b + 2a + c) + |W|(2c') \;=\; 1551.63\lambda \end{aligned}$$

and for $k = 2$ we have

$$\begin{aligned} cost(g, h) &= |U|a + |V|b + |W|c' \;=\; 603.84\lambda \\ cost'(u, u') &= |U|(2a) + |V|(a + b) + |W|(a + c) \;=\; 1294.39\lambda \\ cost'(u, v) &= |U|(a + b) + |V|(2b) + |W|(a + c) \;=\; 1214.39\lambda \\ cost'(u, w) &= |U|(2a) + |V|(a + b) + |W|(2c') \;=\; 1213.93\lambda \end{aligned}$$

Then

$$\begin{aligned} \frac{\min\{cost'(v), cost'(w)\}}{cost(f)} &= \frac{2039}{1014} \;>\; 2.01, \quad \text{and} \\ \frac{\min\{cost'(u, u'), cost'(u, v), cost'(u, w)\}}{cost(g, h)} &= \frac{121393}{60384} \;>\; 2.01. \end{aligned}$$

This implies that inequalities (2), (3) hold with $R = 2.01$ for the modified cost function. But, as we observed earlier, for any facility set F of cardinality $k = 1, 2$, we have $cost'(F) = (1 + \Theta(1/n))cost(F)$. Therefore we can conclude that inequalities (2) and (3) will also hold if we take n large enough, and the lower bound follows. □

The lower bound proof above may seem mysterious, and a reader may wonder how did we discover this specific space and strategy. In fact, we tried to prove an upper bound of 2 for the case when k takes only values 1, 2 and n. In the course of this work, we isolated metric spaces for which we were not able to prove the upper bound – essentially the same spaces as the one in Figure 1. Then, by parametrizing the distances and set cardinalities, with a help of a computer program, we were able to design the lower bound strategy above.

4 A New Upper Bound

In this section we construct an incremental medians sequence with competitive ratio $R = 2 + 4\sqrt{2}$. First, we show that, given a facility set H we can find subsets $F \subseteq G \subseteq H$ of specified sizes and of appropriately small cost. We then use this result to construct our incremental medians sequence.

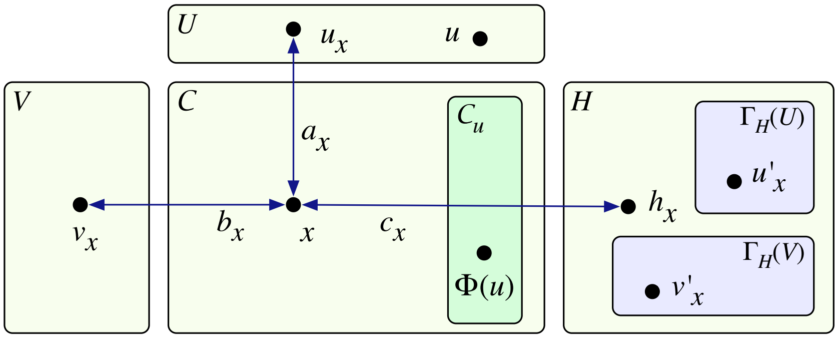

Figure 2: Notation.

4.1 Choosing Two Nested Facility Sets

Let $1 \leq k \leq l \leq m \leq n$. (Recall that $n = |\mathcal{F}|$ is the number of facilities.) Throughout this section we consider three facility sets: H of cardinality m, U of cardinality k, and V of cardinality l. Intuitively, U and V represent optimal $k-$ and $l-$ medians. We use a probabilistic argument to show that there exist two sets F and G, with $|F| = k$, $|G| = l$ and $F \subseteq G \subseteq H$, such that $cost(F)$ and $cost(G)$ are bounded in terms of $cost(U)$, $cost(V)$ and $cost(H)$.

Lemma 2. *Let $1 \leq k \leq l \leq m \leq n$, and let U, V and H be facility sets with $|H| = m$, $|V| = l$ and $|U| = k$. Then there is a set $T \subseteq V$ with $|T| = k$ such that, denoting $\bar{T} = V - T$, we have*

$$cost(\Gamma_H(T)) + cost(\Gamma_H(U \cup \bar{T})) \;\leq\; 2 \cdot cost(H) + 4 \cdot cost(V) + 2 \cdot cost(U). \tag{4}$$

Proof. We use a probabilistic argument, by defining a probability distribution on subsets $T \subseteq V$ and proving that inequality (4) holds in expectation.

Define a random mapping $\Phi : U \to \mathcal{C}$, where $\Phi(u)$ is chosen uniformly from the set $\mathcal{C}_u = \{x \in \mathcal{C} \mid \Gamma_U(x) = u\}$. (If $\mathcal{C}_u = \emptyset$, $\Phi(u)$ is undefined. Alternatively, one can simply remove this u from U, perform the construction for $k-1$ to get a set T with $k-1$ facilities, and then simply add an arbitrary facility to T.) In other words, $\Phi(u)$ is a random customer of u when U is the facility set. Order arbitrarily the elements of V, and for any given Φ define T_Φ as the subset of V that consists of $\Gamma_V(\Phi(U))$ and $k - |\Gamma_V(\Phi(U))|$ smallest elements of V (with respect to the chosen ordering) that are not in $\Gamma_V(\Phi(U))$. Thus $|T_\Phi| = k$.

For each point x in $\mathcal{C}$, let $u_x = \Gamma_U(x)$, $v_x = \Gamma_V(x)$ and $h_x = \Gamma_H(x)$ be the points serving x respectively in U, V and H. The corresponding distances from x are denoted $a_x = d(x, u_x)$, $b_x = d(x, v_x)$ and $c_x = d(x, h_x)$. Let also $u'_x = \Gamma_H(u_x)$ and $v'_x = \Gamma_H(v_x)$. (See Figure 2.)

We now temporarily fix the mapping Φ and a customer $x \in \mathcal{C}$. To simplify notation, we write $T_\Phi = T$ and $u = u_x$. Recall that, for a set F, by $d(x, F) = \min_{f \in F} d_{xf}$ we denote the distance from a point x to the nearest point in a set F. We claim that

$$d(x, \Gamma_H(T)) + d(x, \Gamma_H(U \cup \bar{T})) \;\leq\; a_x + 2b_x + c_x + a_{\Phi(u)} + 2b_{\Phi(u)} + c_{\Phi(u)}. \tag{5}$$

To prove the claim, we consider two cases, for $v_x \in T$ and $v_x \in \bar{T}$.

<u>Case 1</u>: $v_x \in \bar{T}$. This case is illustrated in Figure 3.

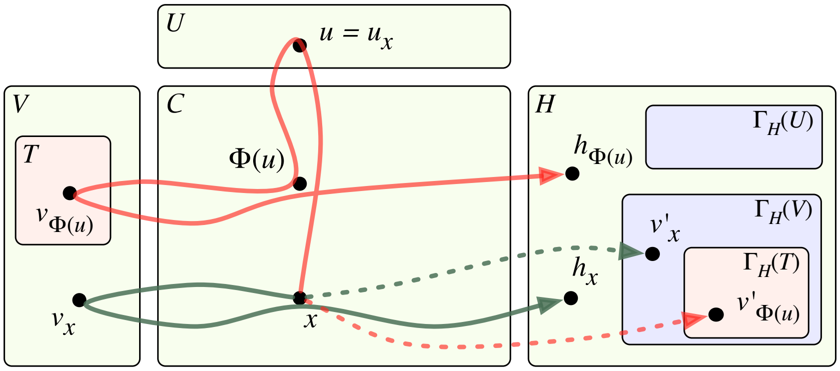

Figure 3: The proof of (5) when $v_x \in \bar{T}$. Dotted lines represent the initial estimates for $d(x, \Gamma_H(T))$ and $d(x, \Gamma_H(U \cup \bar{T}))$ while solid lines show the final estimates.

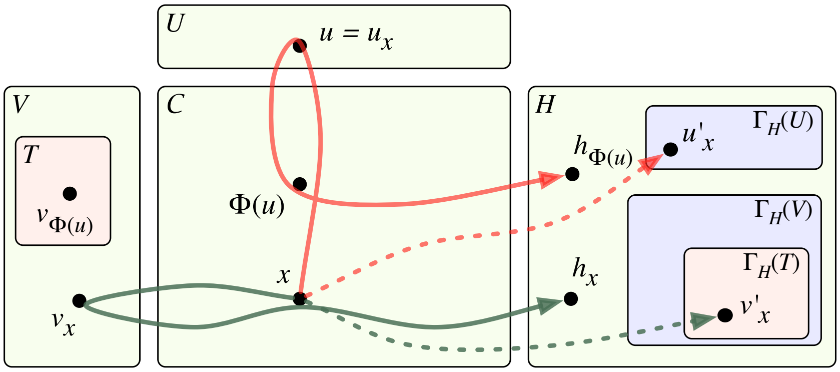

Figure 4: The proof of (5) when $v_x \in \bar{T}$. Dotted lines represent the initial estimates for $d(x, \Gamma_H(T))$ and $d(x, \Gamma_H(U \cup \bar{T}))$, while solid lines show the final estimates.

Since $v'_{\Phi(u)} \in \Gamma_H(T)$, using the definition of $v'_{\Phi(u)}$ and several applications of the triangle inequality, we have $d(x, \Gamma_H(T)) \leq d(x, v'_{\Phi(u)}) \leq a_x + d(u, v_{\Phi(u)}) + d(v_{\Phi(u)}, v'_{\Phi(u)}) \leq a_x + [a_{\Phi(u)} + b_{\Phi(u)}] + d(v_{\Phi(u)}, h_{\Phi(u)}) \leq a_x + a_{\Phi(u)} + 2b_{\Phi(u)} + c_{\Phi(u)}$.

Since $v'_x \in \Gamma_H(U \cup \bar{T})$, using the definition of v'_x and the triangle inequality, $d(x, \Gamma_H(U \cup \bar{T})) \leq d(x, v'_x) \leq b_x + d(v_x, v'_x) \leq b_x + d(v_x, h_x) \leq 2b_x + c_x$.

Combining the two bounds, we get

$$d(x, \Gamma_H(T)) + d(x, \Gamma_H(U \cup \bar{T})) \ \leq \ a_x + 2b_x + c_x + a_{\Phi(u)} + 2b_{\Phi(u)} + c_{\Phi(u)}.$$

Case 2: $v_x \in T$. This case is illustrated in Figure 4.

Since $v'_x \in \Gamma_H(T)$, using the triangle inequality and the definition of v'_x, we have $d(x, \Gamma_H(T)) \leq d(x, v'_x) \leq b_x + d(v_x, v'_x) \leq b_x + d(v_x, h_x) \leq 2b_x + c_x$.

Since $u'_x \in \Gamma_H(U \cup \bar{T})$, using the definition of $u'_x = \Gamma_H(u)$, we have $d(x, \Gamma_H(U \cup \bar{T})) \leq d(x, u'_x) \leq a_x + d(u, u'_x) \leq a_x + d(u, h_{\Phi(u)}) \leq a_x + a_{\Phi(u)} + c_{\Phi(u)}$.

Combining the two bounds we get

$$\begin{aligned} d(x,\Gamma_H(T)) + d(x,\Gamma_H(U\cup\bar{T})) &\leq a_x + 2b_x + c_x + a_{\Phi(u)} + c_{\Phi(u)} \\ &\leq a_x + 2b_x + c_x + a_{\Phi(u)} + 2b_{\Phi(u)} + c_{\Phi(u)}, \end{aligned}$$

completing the proof of inequality (5).

From (5), for a fixed Φ we have

$$\begin{aligned} \mathit{cost}(\Gamma_H(T_\Phi)) + \mathit{cost}(\Gamma_H(U\cup\bar{T}_\Phi)) &\leq \sum_{u\in U}\sum_{x\in\mathcal{C}_u}\left[a_x + 2b_x + c_x + a_{\Phi(u)} + 2b_{\Phi(u)} + c_{\Phi(u)}\right] \\ &\leq \mathit{cost}(H) + 2\cdot\mathit{cost}(V) + \mathit{cost}(U) \\ &\qquad + \sum_{u\in U}|\mathcal{C}_u|\cdot[a_{\Phi(u)} + 2b_{\Phi(u)} + c_{\Phi(u)}]. \end{aligned} \tag{6}$$

For any facility set Z, we have $\mathit{cost}(Z) = \sum_{u\in U}\sum_{x\in\mathcal{C}_u} d(x,Z) = \sum_{u\in U}|\mathcal{C}_u|\cdot \mathrm{Exp}_\Phi[d(\Phi(u),Z)]$, because $\Phi(u)$ is uniformly distributed in $\mathcal{C}_u$. Applying it to $Z = U$, V and H, and using the linearity of expectation, inequality (6) yields

$$\begin{aligned} \mathrm{Exp}_\Phi\left[\mathit{cost}(\Gamma_H(T_\Phi)) + \mathit{cost}(\Gamma_H(U\cup\bar{T}_\Phi))\right] &\leq \mathit{cost}(H) + 2\cdot\mathit{cost}(V) + \mathit{cost}(U) \\ &\qquad + \sum_{u\in U}|\mathcal{C}_u|\cdot\mathrm{Exp}_\Phi\left[a_{\Phi(u)} + 2b_{\Phi(u)} + c_{\Phi(u)}\right] \\ &= 2\cdot\mathit{cost}(H) + 4\cdot\mathit{cost}(V) + 2\cdot\mathit{cost}(U). \end{aligned}$$

This implies that there is a $T = T_\Phi$ that satisfies the lemma. □

Theorem 3. *Let $1 \leq k \leq l \leq m \leq n$. For any facility sets H, U and V with $|U| = k$, $|V| = l$, $|H| = m$, there exist $F \subseteq G \subseteq H$ with $|F| = k$, $|G| = l$ such that*

(i) *$\mathit{cost}(F) \leq \mathit{cost}(H) + 2\cdot\mathit{cost}(U)$ and*

(ii) *$\mathit{cost}(G) \leq \mathit{cost}(H) + 4\cdot\mathit{cost}(V)$.*

Proof. Let $U' = \Gamma_H(U)$ and $V' = \Gamma_H(V)$ be the facilities in H that are closest to those in U and V, respectively. Using the triangle inequality, it is not difficult to show (see [5, 4], for example) that $\mathit{cost}(U') \leq \mathit{cost}(H) + 2\cdot\mathit{cost}(U)$ and $\mathit{cost}(V') \leq \mathit{cost}(H) + 2\cdot\mathit{cost}(V)$.

Let $T \subseteq V$ be the set from Lemma 2. Then either $\mathit{cost}(\Gamma_H(T)) \leq \mathit{cost}(H) + 2\cdot\mathit{cost}(U)$ or $\mathit{cost}(\Gamma_H(U\cup\bar{T})) \leq \mathit{cost}(H) + 4\cdot\mathit{cost}(V)$. In the first case, we take $F = \Gamma_H(T)$ and $G = V'$, and in the second case we take $F = U'$ and $G = \Gamma_H(U\cup\bar{T})$. (If $|F| < k$ or $|G| < l$, we can increase their cardinalities by adding a sufficient number of elements of H while preserving the inclusion $F \subseteq G$.) The theorem then follows from Lemma 2 and the bounds on $\mathit{cost}(U')$ and $\mathit{cost}(V')$. □

4.2 Competitive Incremental Medians

Recall that n is the number of facilities, F^*_j is the optimal j-median and $\mathit{opt}_j = \mathit{cost}(F^*_j)$, for each $j = 1, 2, ..., n$. Our objective is to construct an incremental medians sequence $F_1 \subseteq F_2 \subseteq ... \subseteq F_n$.

The general approach is similar to that in [5, 4]: we construct the sequence backwards, at each step extracting a smaller set of facilities from among those selected earlier. These sets F_j will be constructed only for values of j in a predefined sequence $\{\kappa(a)\}$ of indices, for which the optimal costs increase exponentially with a. For the intermediate values of j, we simply let F_j to be $F_{\kappa(a)}$, where a is the smallest index for which $\kappa(a) \leq j$.

The crucial difference between our method and the previous constructions is in how we extract facilities from $F_{\kappa(a)}$ to form $F_{\kappa(a+1)}$. The algorithms in [5] and [4] select $\kappa(a+1)$ facilities in $F_{\kappa(a)}$ that are closest to those in the optimal set $F^*_{\kappa(a+1)}$. Instead, we use our probabilistic construction from the previous section to simultaneously extract *two* facility sets next in the sequence, namely $F_{\kappa(a+1)}$ and $F_{\kappa(a+2)}$, with Theorem 3 providing an upper bound on their costs.

Construction of incremental medians. Without loss of generality we can assume that $opt_n = 1$, for otherwise we can normalize the instance by dividing all distances by opt_n. (If $opt_n = 0$, instead of n, we can start the process with the largest n' for which $opt_{n'} > 0$.)

We use two parameters $\gamma = 2 + \sqrt{2}/2 \approx 2.71$ and $\lambda = 3\sqrt{2}/2 - 1 \approx 1.16$. We now define a sequence of indices $n = \kappa(0) \geq \kappa(1) \geq ... \geq \kappa(h) = 1$. For $a = 0, 1, ...,$ let

$$\kappa(a) \quad = \quad \begin{cases} \min\{j \mid opt_j \leq \gamma^{a/2}\} & \text{if } a \text{ is even} \\ \min\{j \mid opt_j \leq \lambda\gamma^{(a-1)/2}\} & \text{if } a \text{ is odd} \end{cases}$$

and choose h to be the smallest a for which $\kappa(a) = 1$. Note that we allow some of the elements in the sequence $\{\kappa(a)\}$ to be equal.

We first define facility sets F_j for $j = \kappa(0), \kappa(1), ..., \kappa(h)$. Initially, $F_{\kappa(0)} = \mathcal{F}$, the set of all facilities. Assume that $F_{\kappa(a)}$ has been already defined for some even a, where $0 \leq a \leq h-2$. In Theorem 3 let $m = \kappa(a)$, $H = F_{\kappa(a)}$, $l = \kappa(a+1)$, $k = \kappa(a+2)$, $V = F^*_{\kappa(a+1)}$ and $U = F^*_{\kappa(a+2)}$. We then choose $F_{\kappa(a+2)} \subseteq F_{\kappa(a+1)} \subseteq F_{\kappa(a)}$ such that

$$cost(F_{\kappa(a+1)}) \quad \leq \quad cost(F_{\kappa(a)}) + 4opt_{\kappa(a+1)}, \quad \text{and} \tag{7}$$

$$cost(F_{\kappa(a+2)}) \quad \leq \quad cost(F_{\kappa(a)}) + 2opt_{\kappa(a+2)}. \tag{8}$$

The existence of such sets is guaranteed by Theorem 3; namely take $F_{\kappa(a+1)} = G$ and $F_{\kappa(a+2)} = F$.

It still remains to address the special case when $a = h-1$. In this case, we still can chose a set $F_{\kappa(h)} = F_{\kappa(a+1)}$ satisfying (7), by using $k = l$ in Theorem 3. (Alternatively, we can take $F_{\kappa(a+1)} = \Gamma_H(F^*_{\kappa(a+1)})$, for $H = F_{\kappa(a)}$, as in [5, 4].)

Next, we extend the sequence to other values of j. If $\kappa(a+1) < j < \kappa(a)$, we simply let $F_j = F_{\kappa(a+1)}$. This completes the construction.

Theorem 4. *The incremental sequence $\{F_j\}$ constructed above is R-competitive, where $R = 2 + 4\sqrt{2} \approx 7.656$.*

Proof. For each $j = 1, ..., n$, denote $cost_j = cost(F_j)$. Using the bounds (7), (8), and the definition of the sequence $\{\kappa(a)\}$, each value $cost_{\kappa(a)}$ can be estimated as follows: if a is even, then $cost_{\kappa(a)} \leq 2\sum_{b=0}^{a/2} opt_{\kappa(2b)} \leq 2\sum_{b=0}^{a/2} \gamma^b$, and if a is odd then $cost_{\kappa(a)} \leq cost_{\kappa(a-1)} + 4opt_{\kappa(a)} \leq 2\sum_{b=0}^{(a-1)/2} \gamma^b + 4\lambda\gamma^{(a-1)/2}$. Summing up the geometric sequences, we thus get

$$cost_{\kappa(a)} \quad \leq \quad \begin{cases} \dfrac{2\gamma^{a/2+1}}{\gamma - 1} & \text{if } a \text{ is even} \\ \dfrac{2\gamma^{(a-1)/2+1}}{\gamma - 1} + 4\lambda\gamma^{(a-1)/2} & \text{if } a \text{ is odd} \end{cases}$$

Fix some number of facilities j, and choose a such that $\kappa(a+1) \leq j < \kappa(a)$. We want to show that $cost_j \leq R \cdot opt_j$. By the construction, $F_j = F_{\kappa(a+1)}$, so $cost_j = cost_{\kappa(a+1)}$. We have two cases.

Suppose first that a is even. By the choice of j and the definition of $\kappa(a)$, we get $opt_j > \gamma^{a/2}$. Since $cost_j = cost_{\kappa(a+1)} \leq 2\gamma^{a/2+1}/(\gamma-1) + 4\lambda\gamma^{a/2}$, the ratio is

$$\frac{cost_j}{opt_j} \;\leq\; \frac{2\gamma}{\gamma-1} + 4\lambda \;=\; R.$$

If a is odd, then by the choice of j and the definition of $\kappa(a)$, we get $opt_j > \lambda\gamma^{(a-1)/2}$. Since $cost_j = cost_{\kappa(a+1)} \leq 2\gamma^{(a+1)/2+1}/(\gamma-1)$, the ratio is

$$\frac{cost_j}{opt_j} \;\leq\; \frac{2\gamma^2}{(\gamma-1)\lambda} \;=\; R,$$

completing the proof.

□

5 2-Competitive Incremental Medians for Equable Instances

In this section we present a construction of a 2-competitive incremental medians sequence for a special case where, for any fixed value of k each customer has the same distance to the optimal k-median.

More formally, we consider the following setting. Suppose $(\mathcal{F}, \mathcal{C})$ is an instance of the medians problem with $\mathcal{C} \subseteq \mathcal{F}$ and $|\mathcal{C}| = m$. For each $k = 1, 2, ..., m$ we are also given an "adversary" k-median F^*_k such that $d(x, F^*_k) = \delta_k$ for all $x \in \mathcal{C}$, where $\delta_1 > \delta_2 > ... > \delta_m$. Thus $cost(F^*_k) = m\delta_k$ for all k. We will refer to this instance as an *equable instance.* We prove that there is an incremental medians sequence $F_1 \subseteq F_2 \subseteq ... \subseteq F_m$ that is 2-competitive against the adversary medians[1], that is $cost(F_k) \leq 2m\delta_k$ for all $k = 1, 2, ..., m$.

The motivation for considering this version is two-fold. First, the original lower bound of 2 [6, 7], as well as most of our own attempts to improve it, were based on equable distances. (The use of such instances is natural, because their symmetry greatly reduces the complexity of reasoning about an online algorithm's behavior.) Our result shows that this approach will not work. Second, it also shows that a ratio lower than that in Section 4 can be achieved if, for each k, the distances between all customers and their facility in F^*_k are sufficiently close to each other. Thus hard instances are those where, for some values of k, the distances between customers and their facilities in F^*_k vary greatly.

Our method in this section is different from previous constructions of incremental medians, including the one from Section 4. Unlike in these previous approaches, we construct the sequence $F_1, F_2, ..., F_m$ *forward*, maintaining an invariant ensuring that we not only do well at step k, but also that we make good progress towards obtaining a low-cost l-median for all $l > k$.

Intuition. We start with a simple, although not quite correct, construction, and later we will explain how to modify it to make it work. Imagine that we can order the customers $x_1, x_2, ..., x_m$ such that, for each k, the first k customers $x_1, x_2, ..., x_k$ are served by different facilities in F^*_k. For

[1] It is convenient here to consider this more general setting of adversary medians rather than the true optimal medians, because optimal medians with the desired property would not exist for the values of $k > m/2$.

any k define $F_k = \{x_1, x_2, ..., x_k\}$. We claim that in this case we have $\mathit{cost}(F_k) \le 2m\delta_k$. Indeed, if $x \in \mathcal{C}$ is any customer, choose the point x_j, $j \le k$, that is served by the same facility f in F^*_k as x. This x_j must exist by the assumption about the sequence $\{x_i\}$. Then $d(x, x_j) \le d(x, f) + d(f, x_j) = 2\delta_k$, and the bound on $\mathit{cost}(F_k)$ follows.

The problem with the argument above is that the sequence $x_1, x_2, ..., x_m$ with the required property may not exist. By relaxing appropriately the condition on the x_k's, we obtain the construction detailed below.

Incremental spanners. We introduce first an auxiliary combinatorial construction. Suppose that for each $k = 1, 2, ..., m$ we have a family $\mathcal{S}_k \subseteq 2^{\mathcal{C}}$ of k sets that forms a partition of $\mathcal{C}$, that is, all sets in $\mathcal{S}_k$ are disjoint and $\bigcup_{A \in \mathcal{S}_k} A = \mathcal{C}$. For a set $X \subseteq \mathcal{C}$, define its *k-span* as

$$\mathit{Span}_k(X) \;=\; \bigcup \{A \in \mathcal{S}_i \mid i \ge k \text{ and } A \cap X \ne \emptyset\}.$$

Note that $X \subseteq \mathit{Span}_k(X)$ for all k, and that $\mathit{Span}_k(X) \subseteq \mathit{Span}_j(X)$ for all $j \le k$. A set $X \subseteq \mathcal{C}$ is called a *k-spanner* if $\mathit{Span}_k(X) = \mathcal{C}$. By the earlier observation, if X is a k-spanner then it is also a j-spanner for any $j < k$.

A sequence $X_1 \subseteq X_2 \subseteq ... \subseteq X_m$ is called an *incremental spanner* if for each $k = 1, 2, ..., m$, $|X_k| = k$ and X_k is a k-spanner. We now show how to construct an incremental spanner.

For $X \subseteq \mathcal{C}$ and any $j = 1, 2, ..., m$, let $\mathit{setscov}_j(X)$ be the collection of sets in $\mathcal{S}_j$ covered by the j-span of X, that is

$$\mathit{setscov}_j(X) \;=\; \{A \in \mathcal{S}_j \mid A \subseteq \mathit{Span}_j(X)\}.$$

Note that $|\mathit{setscov}_j(X)| = j$ if and only if X is a j-spanner, because $\mathcal{S}_j$ is a partition of $\mathcal{C}$.

We will construct sets $\emptyset = X_0 \subseteq X_1 \subseteq ... \subseteq X_m$ so that, for each $k = 0, 1, 2, ..., m$, we will have $|X_k| = k$ and the following invariant will hold:

$$|\mathit{setscov}_j(X_k)| \;\ge\; k, \quad \text{for all } j = k, k+1, ..., m. \tag{9}$$

Initially, for $k = 0$, we set $X_0 = \emptyset$, and (9) holds trivially. Suppose we have $X_0, X_1, ..., X_{k'}$, for some $k' < m$ and that (9) holds for $k = 0, 1, ..., k'$. This implies, in particular, that $|\mathit{setscov}_{k'}(X_{k'})| = k'$, that is, $X_{k'}$ is a k'-spanner. Thus $X_{k'}$ is also a k-spanner for all $k \le k'$. Let l be the minimum index for which $X_{k'}$ is *not* an l-spanner, that is $\mathcal{C} - \mathit{Span}_l(X_{k'}) \ne \emptyset$. By the choice of l, we have $l > k'$. Pick any $x \in \mathcal{C} - \mathit{Span}_l(X_{k'})$ and take $X_{k'+1} = X_{k'} \cup \{x\}$. Clearly, $|X_{k'+1}| = k' + 1$, because $x \notin X_{k'}$.

We now show that (9) holds for $k = k'+1$. By the choice of l, for $j = k'+1, k'+2, ..., l-1$, $X_{k'}$ is a j-spanner. Therefore, for these values of j, $X_{k'+1}$ is a j-spanner as well, and thus $|\mathit{setscov}_j(X_{k'+1})| = j \ge k' + 1$. We thus have that (9) holds for $j = k' + 1, k' + 2, ..., l - 1$ and $k = k' + 1$. Consider any $j \ge l \ge k' + 1$. Let $A \in \mathcal{S}_j$ be the set for which $x \in A$. Since $x \in \mathcal{C} - \mathit{Span}_l(X_{k'}) \subseteq \mathcal{C} - \mathit{Span}_j(X_{k'})$, we have $A \notin \mathit{setscov}_j(X_{k'})$. But now $x \in X_{k'+1}$, so $A \in \mathit{setscov}_j(X_{k'+1})$ and, by induction, we get $|\mathit{setscov}_j(X_{k'+1})| \ge |\mathit{setscov}_j(X_{k'})| + 1 \ge k' + 1$. This completes the proof that our construction preserves invariant (9).

By (9), for each k we have $|\mathit{setscov}_k(X_k)| \ge k$, and thus X_k is a k-spanner. We can conclude then that $X_1, X_2, ..., X_m$ is an incremental spanner.

Incremental medians. We now show how to use incremental spanners to construct incremental medians. For $k = 1, 2, ..., m$, assign each customer $x \in \mathcal{C}$ to its closest facility $f \in F^*_k$ (that is,

$d_{xf} = \delta_k$), breaking ties arbitrarily. Define C_k^f to be the set of customers assigned to f, and let $\mathcal{S}_k = \{C_k^f \mid f \in F_k^*\}$. Then each $\mathcal{S}_k$ contains k sets and forms a partition of $\mathcal{C}$. As we showed above, for these partitions $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_m$ there exists an incremental spanner $F_1, F_2, ..., F_m$.

We claim that $F_1, F_2, ..., F_m$ is 2-competitive against the adversary medians. Consider some fixed k. Since F_k is a k-spanner, for each customer $x \in \mathcal{C}$ there is $i \geq k$ and $f \in F_i^*$ such that $x \in C_i^f$ and $C_i^f \cap F_k \neq \emptyset$. Choose any $y \in C_i^f \cap F_k$. Then $d(x, F_k) \leq d_{xy} \leq d_{xf} + d_{yf} = 2\delta_i \leq 2\delta_k$. This implies that $cost(F_k) \leq 2m\delta_k$, and the claim follows.

Summarizing, we obtain the following result:

Theorem 5. *Any equable instance of the medians problem has an incremental medians sequence that is* 2*-competitive against the adversary medians.*

6 Final Comments

We improved both the lower and upper bounds for incremental medians, from 2 to 2.01 and from 8 to $2 + 4\sqrt{2} \approx 7.656$, respectively, thus proving that neither 2 nor 8 are the "right" bounds for this problem. (By optimizing the parameters in Section 3 it is possible to improve the lower bound slightly, to about 2.01053.) In addition to its own independent interest, closing or significantly reducing the remaining gap would shed more light on the computational hardness of approximating incremental medians, as it would show to what degree the difficulty of the problem can be attributed to non-existence of incremental median sequences with small competitive ratios.

The expected values in the proof of Lemma 2 can be computed in polynomial-time, and thus our probabilistic construction in that proof can be de-randomized using the method of conditional expectations. This does not necessarily lead to a polynomial-time construction of incremental medians, since our construction in Section 4 requires the knowledge of optimal k-medians for all values of k. It is possible, however, that our method from Lemma 2 can be combined with the approach from [5] to obtain a ratio below 16 in polynomial time. Since the potential improvement, if possible at all, appears to be minor, we did not pursue this direction of research.

We believe that some of the ideas in the paper can be used to prove even better bounds. In the upper bound proof in Section 4 we construct our sequence backwards, starting with all facilities, and gradually extracting smaller and smaller facility sets, two at a time. By extending the probabilistic construction to more than two steps at a time, we should be able to get a better bound. Even our two-step method still might have room for improvement, as the two choices for F and G considered in the proof of Theorem 3 are not "balanced", that is, the bounds on the cost of F and G in the two cases are not the same. Also, our construction of a 2-competitive incremental medians sequence for equable spaces is very different from previous constructions and we believe that its basic idea will be useful in improving the upper bound for general spaces.

Our lower bound argument uses only three steps, for $k = 1, 2, n$. It should be possible to improve our bound by using either $k > 2$ as the intermediate number of facilities or more (perhaps an unbounded number of) steps. Both ideas lead to difficulties that we were not able to overcome at this time. In a three-step strategy using $k = 1, k', n$ with $k' > 2$, an algorithm can place facilities $2, .., k'$ optimally (given the choice of the first facility), and thus increasing k' seems only to help the algorithm. A strategy that uses additional steps leads to a different problem. Average costs for the customers must decrease with k, and thus introducing additional steps creates shortcuts via optimal k'-medians for large k', reducing the algorithm's cost for small values of k.

The result from Section 5 may also be useful for lower bound proofs, as it shows that in "hard" instances, for a fixed k, the optimal customers' costs should be significantly different.

Acknowledgments. We would like to thank the anonymous referees for useful comments and for pointing out several mistakes in the earlier draft of this paper.

References

[1] Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristic for k-median and facility location problems. In *Proc. 33rd Symp. Theory of Computing (STOC)*, pages 21–29. ACM, 2001.

[2] Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for k-median and facility location problems. *SIAM Journal on Computing*, 33(3):544–562, 2004.

[3] Marek Chrobak and Claire Kenyon. Competitiveness via doubling. *SIGACT News*, pages 115–126, 2006.

[4] Marek Chrobak, Claire Kenyon, John Noga, and Neal Young. Online medians via online bidding. In *Proc. 7th Latin American Theoretical Informatics Symp. (LATIN)*, volume 3887 of *Lecture Notes in Comput. Sci.*, pages 311–322, 2006.

[5] Guolong Lin, Chandrashekha Nagarajan, Rajmohan Rajamaran, and David P. Williamson. A general approach for incremental approximation and hierarchical clustering. In *Proc. 17th Symp. on Discrete Algorithms (SODA)*, pages 1147–1156, 2006.

[6] Ramgopal R. Mettu and C. Greg Plaxton. The online median problem. In *Proc. 41st Symp. Foundations of Computer Science (FOCS)*, pages 339–348. IEEE, 2000.

[7] Ramgopal R. Mettu and C. Greg Plaxton. The online median problem. *SIAM J. Comput.*, 32:816–832, 2003.