One-Dimensional r-Gathering under Uncertainty

Shareef Ahmed[1], Shin-ichi Nakano[2], and Md. Saidur Rahman[1]

[1] Graph Drawing and Information Visualization Laboratory,
Department of Computer Science and Engineering,
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
{shareefahmed,saidurrahman}@cse.buet.ac.bd
[2] Gunma University, Kiryu 376-8515, Japan
nakano@cs.gunma-u.ac.jp

Abstract. Let C be a set of n customers and F be a set of m facilities. An r-gathering of C is an assignment of each customer $c \in C$ to a facility $f \in F$ such that each facility has zero or at least r customers. The r-gathering problem asks to find an r-gathering that minimizes the maximum distance between a customer and its facility. In this paper we study the r-gathering problem when the customers and the facilities are on a line, and each customer location is uncertain. We show that, the r-gathering problem can be solved in $O(nk+mn\log n+(m+n\log k+n\log n+nr^{\frac{n}{r}})\log mn)$ and $O(mn\log n+(n\log n+m)\log mn)$ time when the customers and the facilities are on a line, and the customer locations are given by piecewise uniform functions of at most $k+1$ pieces and "well-separated" uniform distribution functions, respectively.

Keywords: r-Gathering, Facility location problem

1 Introduction

The facility location problem and many of its variants are well studied [7]. In this paper we study a relatively new variant of the facility location problem, called the r-gathering problem [6].

Let C be a set of n customers and F be a set of m facilities, $d(c, f)$ be the distance between $c \in C$ and $f \in F$. An r-gathering of C to F is an assignment A of C to F such that each facility has at least r or zero customers assigned to it. The cost of an r-gathering is $\max_{c \in C}\{d(c, A(c))\}$ which is the maximum distance between a customer and its facility. The r-gathering problem asks to find an assignment of C to F having the minimum cost [6]. This problem is also known as the min-max r-gathering problem. The other version of the problem is known as the min-sum r-gathering problem which asks to find an assignment which minimizes $\sum_{c\in C} d(c, A(c))$ [8, 11]. In this paper we consider the min-max r-gathering problem and we use the term r-gathering problem to refer the min-max version.

Assume we wish to set up emergency shelters for residents C living on a locality so that each shelter can accommodate at least r residents. We also wish to locate the shelters so that evacuation time span can be minimized. A set F of possible locations for shelters is also given. This scenario can be modeled by the r-gathering problem. In this case, an r-gathering corresponds to an assignment of residents to shelters so that each "open" shelter serves at least r residents and the r-gathering problem finds the r-gathering minimizing the evacuation time.

For the r-gathering problem a 3-approximation algorithm is known and it is proved that the problem cannot be approximated within a factor less than 3 for $r > 3$ unless $P = NP$ [6]. Recently, the problem is considered in a setting where all the customers and facilities are lying on a line. An $O((n+m)\log(n+m))$ time algorithm [5], an $O(n+m\log^2 r+m\log m)$ time algorithm [9], an $O(n + r^2m)$ time algorithm [12], and an $O(n + m)$ time algorithm [13] are known when all the customers and facilities are on a line. Ahmed *et al.* gave an $O(n+m+d^2r^2(d+\log m)+(r+1)^d 2^d(r+d)d)$ time algorithm for the r-gathering problem when the customers and facilities are on a star [4].

In this paper, we consider the r-gathering problem when the customer and the facilities are on a line, and the customer locations are uncertain. Study of different problems under uncertain settings become much popular recently. Uncertainty in data usually occurs because of noise in measured data, sampling inaccuracy, limitation of resources, etc. Hence uncertainty is ubiquitous in practice and managing the uncertain data has gained much attention [1–3, 15]. Different variants of the facility location problem have also been investigated under uncertain settings. Setting up a facility is costly and each facility is supposed to serve for a long period of time. On the other hand existence, location and demand of a client can change over time. Thus it is important to set up facilities by keeping the uncertainty in mind. For the detailed state of the art of uncertain facility location problem, we refer the survey of Snyder [14]. There are two models for uncertainty: one is existential model [10, 18] and the other is locational model [1, 2, 16]. In the existential model, the existence of each point is uncertain. Thus each point has a specific location and there is a probability for the existence of each point. In the locational model each point is certain to exist, but its position in uncertain and defined by a probability density function. In this paper we consider the locational model of uncertainty. For customer locations, we consider two probability density functions: piecewise uniform function (histogram) and "well-separated" uniform distribution function.

When the customer and facility locations are deterministic and on a line, there is an optimal r-gathering where the customers assigned to each facility are consecutive on the line [12]. However, when the customer locations are uncertain, finding a suitable ordering of the customers is difficult. In this paper we give an $O(nk+mn\log n+(m+n\log k+n\log n+nr^{\frac{n}{r}})\log mn)$ time algorithm for the one-dimensional r-gathering problem when the customer locations are given by piecewise uniform functions of at most $k+1$ pieces, and an $O(mn\log n+(n\log n+m)\log mn)$ time algorithm for the one-dimensional r-gathering problem when the customer locations are given by well-separated uniform distributions.

The rest of the paper is organized as follows. In Section 2, we define the uncertain r-gathering problem and provide definitions of basic terminologies. In Section 3, we give algorithms for uncertain r-gathering problem when customer locations are specified by piecewise uniform functions and "well-separated" uniform distribution functions. Finally we conclude in Section 4.

2 Preliminaries

In this section we define the uncertain r-gathering problem and relevant terminologies.

Let $F=\{f_1,f_2,\cdots,f_m\}$ be a set of m facilities, and $\mathcal{C}=\{C_1,C_2,\cdots,C_n\}$ be a set of n customers where each C_i is a random variable. The probability density function (PDF) associated with customer C_i is denoted by $g_i(x)$. The expected distance between a facility f_j and an uncertain customer C_i, denoted by $E[d(C_i,f_j)]$, is $\int_{-\infty}^{\infty} d(x,f_j)g_i(x)dx$. An *$r$-gathering* A of $\mathcal{C}$ to F is an assignment $A:\mathcal{C}\rightarrow F$ such that each facility serves zero or at least r customers. A facility having one or more customers is called an *open facility*. $A(C)$ denotes the facility to which a customer C is assigned in an assignment A. The cost of a facility is the maximum expected distance between the facility and its customers if the facility is open, and zero otherwise. The cost of an r-gathering is the maximum cost among all the facilities. The *uncertain r-gathering problem* asks to find an r-gathering with minimum cost. Note that, the uncertain r-gathering problem is NP-Hard, since it contains the deterministic version as a special case.

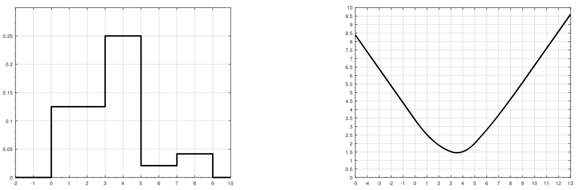
3 One-dimensional Uncertain r-Gathering Problem

In this section we give two algorithms for the uncertain r-gathering problem on a line.

Let $\mathcal{C} = \{C_1, C_2, \cdots, C_n\}$ be a set of n uncertain customer on a horizontal line where each customer C_i is specified by its PDF $g_i : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$, and $F = \{f_1, f_2, \cdots, f_m\}$ be a set of m facilities on the horizontal line. We consider the facilities are ordered from left to right. An r-gathering of $\mathcal{C}$ to F is an assignment $A : \mathcal{C} \to F$ such that each facility serves zero or at least r customers. The uncertain r-gathering problem asks to find an r-gathering such that the maximum among the expected distances between a customer to the assigned facility is minimum.

3.1 Histogram

In this section we give an algorithm for the uncertain r-gathering problem when each customer location is specified by a piecewise uniform function, i.e., a histogram.



(a) (b)

Fig. 1. (a) Illustration of a histogram and (b) corresponding function of expected distance.

We consider the PDF of each customer C_i is defined as a piecewise uniform function g_i, i.e., a histogram. The PDF of each uncertain customer is independent. We consider histogram model since it can be used to approximate any PDF [1]. The histogram model is considered by Wang and Zhang [17] for the uncertain k-center problem on a line. Each g_i consists of at most $k+1$ pieces where each piece is a uniform function. Each customer C_i has $k+2$ points $x_{i0}, x_{i1}, \cdots, x_{i(k+1)}$, where $x_{i0} < x_{i1} < \cdots < x_{i(k+1)}$, and $k+1$ values $y_{i0}, y_{i1}, \cdots, y_{ik}$ such that $g_i(x) = y_{ij}$ if $x_{ij} \leq x < x_{i(j+1)}$. We consider $x_{i0} = -\infty$, $x_{i(k+1)} = \infty$, $y_0 = 0$, and $y_k = 0$. Figure 1(a) illustrates a histogram of 6 pieces. The expected distance $E[d(p, C_i)]$ from a point p to C_i is defined as follows.

$$E[d(p, C_i)] = \int_{-\infty}^{\infty} g_i(x)|x - p|dx$$

A function $h : \mathbb{R} \to \mathbb{R}$ is called a *unimodal function* if there is a point p such that $h(x)$ is monotonically decreasing in $(-\infty, p]$ and monotonically increasing in $[p, \infty)$. Wang and Zhang gave the following lemma [17].

Lemma 1 ([17]). *Let C_i be an uncertain point on a line which is specified by a histogram of $k+1$ pieces. Then the function $E[d(p, C_i)]$ for $p \in \mathbb{R}$ is a unimodal function consisting of a parabola in each interval $[x_{ij}, x_{i(j+1)})$. Furthermore the function $E[d(p, C_i)]$ can be explicitly computed in $O(k)$ time.*

Outline of the Proof. Without loss of generality, assume that $x_{it} \leq p \leq x_{i(t+1)}$. Then the function $E[d(p, C_i)]$ can be written as follows [17].

$$E[d(p,C_i)] = y_{it}p^2 + \left[\sum_{j=0}^{t-1} y_{ij}\left(x_{i(j+1)} - x_{ij}\right) - \sum_{j=t+1}^{k} y_{ij}\left(x_{i(j+1)} - x_{ij}\right) - y_{it}(x_{it} + x_{i(t+1)})\right] p$$
$$+ \frac{1}{2}\left[\sum_{j=t+1}^{k} y_{ij}\left(x_{i(j+1)}^2 - x_{ij}^2\right) - \sum_{j=0}^{t-1} y_{ij}\left(x_{i(j+1)}^2 - x_{ij}^2\right) + y_{it}(x_{it}^2 + x_{i(t+1)}^2)\right] \quad (1)$$

Thus we can write $E[d(p, C_i)]$ as $a_{i1}(t)p^2 + a_{i2}(t)p + a_{i3}$ where each of $a_{i1}(t), a_{i2}(t), a_{i3}(t)$ depends on t satisfying $x_{it} \leq p \leq x_{i(t+1)}$. Note that if $y_{it} = 0$ then the function $E[d(p, C_i)]$ is a straight line in the interval $[x_{it}, x_{i(t+1)})$ which we consider as a special parabola. Figure 1(b) illustrates the $E[d(p, C_i)]$ function for the histogram in Figure 1(a). We can compute the co-efficients $a_{i1}(j)$ for all j in $O(k)$ time. Moreover, the summation terms in $a_{i2}(j)$ and $a_{i3}(j)$ for all j can be computed in $O(k)$ time in total. Thus for all j, we can compute the $a_{i2}(j)$ and $a_{i3}(j)$ in $O(k)$ time. Hence the function $E[d(p, C_i)]$ can be computed explicitly in $O(k)$ time. □

We now give the following lemma.

Lemma 2. *Let C_i be an uncertain point on a line which is specified by a histogram of $k + 1$ pieces, and $F = \{f_1, f_2, \cdots, f_m\}$ be a set of m facilities on the line. We can compute the expected distances between all facilities and the uncertain point in $O(m+k)$ time. Furthermore the expected distances between the facilities and the uncertain point can be sorted in $O(m)$ time.*

Proof. We first precompute the co-efficients $a_{i1}(j), a_{i2}(j), a_{i3}(j)$ of function $E[d(p, C_i)]$ for all j in $O(k)$ time by Lemma 1. With the precomputed function $E[d(p, C_i)]$, the expected distance between the uncertain point and a facility f_u can be computed in $O(\log(k))$ time using binary search to find the $[x_{it}, x_{i(t+1)})$ where f_u is located. Thus the expected distance between all facilities and the uncertain point can be computed in $O(m \log k)$ time. However, we can improve the running time to $O(m + k)$ performing a plane sweep from left to right. We take the facilities from left to right, determine the corresponding interval $[x_{ij}, x_{i(j+1)})$, and compute the expected distance. Since both the facilities and the $x_{i1}, x_{i2}, \cdots, x_{ik}$ are ordered from left to right, the search for the interval in which f_u is located can start from the interval in which f_{u-1} is located. Hence each x_{ij} will be considered once. Thus the total running time is $O(m + k)$. We now show that the sorted list of the expected distances between the facilities and the uncertain point can be constructed in $O(m + k)$ time. Since $E[d(p, C_i)]$ is a unimodal function, there is a facility f_u such that $E[d(f_{v-1}, C_i)] \geq E[d(f_v, C_i)]$ for any $1 < v \leq u$, and $E[d(f_v, C_i)] \leq E[d(f_{v+1}, C_i)]$ for any $u \leq v < m$. Thus we have a descending list of expected distances for $f_1, f_2, \cdots, f_u$ and ascending list of expected distances for $f_{u+1}, f_{u+2}, \cdots, f_m$. We can merge these two lists into an ascending list of expected distances in $O(m)$ time. □

Corollary 1. *Let $\mathcal{C} = \{C_1, C_2, \cdots, C_n\}$ be set of n uncertain customers on a line each of which is specified by a histogram of $k + 1$ pieces, and $F = \{f_1, f_2, \cdots, f_m\}$ be a set of m facilities on the line. The expected distances between all pair of uncertain customers and facilities can be computed and sorted in $O(nk + mn \log n)$ time.*

Proof. By Lemma 2, we can compute n sorted list of expected distances between customers and facilities in $O(nk + mn)$ time. The n sorted lists can be merged into a single list using min-heap in $O(mn \log n)$ time. □

We first consider the decision version of the uncertain r-gathering problem on a line. Given a set of uncertain customers $\mathcal{C}$, a set of facilities F on a line, and a number b, the decision uncertain

r-gathering problem asks to determine whether there is an r-gathering A of $\mathcal{C}$ to F such that $E[d(C, A(C))] \leq b$ for each $C \in \mathcal{C}$. The following lemma is known [17].

Lemma 3 ([17]). *Let C be an uncertain point on a line which is specified by a histogram of $k+1$ pieces, and b is a number. Then the points p for which $E[d(C,p)] \leq b$ holds form an interval on the line.*

We call the interval which admits $E[d(C,p)] \leq b$ for customer C a (C,b)-interval and denote the interval by $[s_b(C), t_b(C)]$. Furthermore in any r-gathering A with cost at most b, $A(C)$ is in $[s_b(C), t_b(C)]$. Thus to find whether there is an r-gathering satisfying $E[d(C,p)] \leq b$ for each customer C, it is sufficient to solve the following problem. Given a set of facilities F on a line and a set of customers $\mathcal{C}$ where each customer $C \in \mathcal{C}$ has an interval $[s(C), t(C)]$ on the line, the *interval r-gathering problem* asks to determine whether there is an r-gathering A such that each facility $f \in F$ serves zero or at least r customers and for each customer $C \in \mathcal{C}$, $s(C) \leq A(C) \leq t(C)$ holds.

We now give an algorithm for the interval r-gathering problem. Let $F = \{f_1, f_2, \cdots, f_m\}$ be a set of facilities and $\mathcal{C} = \{C_1, C_2, \cdots, C_n\}$ be a set of customers on a line where each customer C_i has an interval $I_i = [s(C_i), t(C_i)]$. An interval I_i is called the *leftmost interval* if for each $C_j \neq C_i$, $t(C_i) \leq t(C_j)$ holds, and the customer C_i is called the *leftmost customer*. A facility f_u is called the *preceding facility* of C_i if $s(C_i) \leq f_u \leq t(C_i)$ and there is no facility f_v such that $f_u < f_v \leq t(C_i)$. Similarly a facility f_u is called the *following facility* of C_i if $s(C_i) \leq f_u \leq t(C_i)$ and there is no facility f_v such that $s(C_i) \leq f_v < f_u$. We call a customer C_j a *right neighbor* of C_i if $t(C_j) \geq t(C_i)$ and $s(C_j) \leq t(C_i)$.

Let $F = \{f_1, f_2, \cdots, f_m\}$ be a set of facilities and $\mathcal{C} = \{C_1, C_2, \cdots, C_n\}$ be a set of customers on a line where each customer C_i has an interval I_i. Let C_i be the leftmost customer, f_u be the preceding facility of C_i, and $\mathcal{C}_u$ be the set of customers containing f_u in their intervals. We now have the following two lemmas.

Lemma 4. *If there is an interval r-gathering of $\mathcal{C}$ to F, then there is an interval r-gathering with the leftmost open facility f_u. Furthermore, the customers assigned to f_u have consecutive right end-points in $\mathcal{C}_u$ including C_i.*

Proof. We first prove that there is an interval r-gathering with the leftmost open facility f_u. Assume for a contradiction that there is no interval r-gathering with the leftmost open facility f_u. Let A be an interval r-gathering with the leftmost open facility $f_v \neq f_u$. We can observe that $f_v \leq f_u$, since in each interval r-gathering C_i is assigned to a facility within the interval I_i and f_u is the preceding facility of C_i. Let $\mathcal{C}'_v$ be the set of customers assigned to f_v in A. For any customer C_j in $\mathcal{C}'_v$, we have $s(C_j) \leq f_v \leq f_u \leq t(C_i) \leq t(C_j)$, since I_i is the leftmost interval. We now derive a new interval r-gathering by reassigning the customers $\mathcal{C}'_v$ to f_u. A contradiction.

We now prove that the customers assigned to f_u have consecutive right end-points in $\mathcal{C}_u$. We call a pair $C_j, C_k \in \mathcal{C}_u$ a reverse pair if $t(C_j) < t(C_k)$, C_k assigned to f_u, and C_j assigned to $f_v > f_u$. Assume for a contradiction that there is no interval r-gathering where the customers assigned to f_u have consecutive right end-points in $\mathcal{C}_u$. Let A' be an interval r-gathering with minimum number of reverse pairs but the number is not zero. Let C_j, C_k be a reverse pair in A' where $t(C_j) < t(C_k)$, and C_j is assigned to facility f_w, and C_k is assigned to f_u. Since $t(C_k) > t(C_j)$ and $f_w \geq f_u$, we get $s(C_k) \leq f_w \leq t(C_k)$. We now derive a new interval r-gathering with less reverse pairs by reassigning C_j to f_u and C_k to f_w, a contradiction. □

Lemma 5. *Let C_j be the leftmost customer in $\mathcal{C} \setminus \mathcal{C}_u$, and $\mathcal{C}'_u \subseteq \mathcal{C}_u$ be the customers such that for each $C \in \mathcal{C}'_u$, $t(C) < t(C_j)$. If there is an interval r-gathering, then there is an interval r-gathering satisfying one of the following.*

(a) If $|\mathcal{C}'_u| < r$, then the customers assigned to f_u are the r leftmost customers in $\mathcal{C}_u$.
(b) If $|\mathcal{C}'_u| \geq r$, then $\max\{|\mathcal{C}'_u| - r + 1, r\}$ leftmost customers of $\mathcal{C}'_u$ are assigned to f_u (possibly with more customers).

Proof. (a) By Lemma 4, the customers assigned to f_u are consecutive in $\mathcal{C}_u$. Thus the leftmost r customers $\mathcal{C}^l_u$ in $\mathcal{C}_u$ are assigned to f_u. We now prove that there is an interval r-gathering where no customer in $\mathcal{C}_u \setminus \mathcal{C}^l_u$ is assigned to f_u. Assume for a contradiction that in every interval r-gathering there are some customers in $\mathcal{C}_u \setminus \mathcal{C}^l_u$ which are assigned to f_u. Let A be an interval r-gathering where the number of customers in $\mathcal{C}_u \setminus \mathcal{C}^l_u$ assigned to f_u is minimum, and C_k be a customer in $\mathcal{C}_u \setminus \mathcal{C}^l_u$ which is assigned to f_u. Since $|\mathcal{C}'_u| < r$, we get $t(C_k) > t(C_j)$. Let C_j is assigned to f_v in A. We now derive a new r-gathering by reassigning C_k to f_v, a contradiction.
(b) We first consider $r \leq |\mathcal{C}'_u| < 2r$. In this case $\max\{|\mathcal{C}'_u| - r + 1, r\} = r$. Hence by Lemma 4 the leftmost r customers in $\mathcal{C}_u$ are assigned to f_u.
We now consider $|\mathcal{C}'_u| \geq 2r$. In this case, $\max\{|\mathcal{C}'_u| - r + 1, r\} = |\mathcal{C}'_u| - r + 1$. Let $\mathcal{C}''_u$ be the leftmost $|\mathcal{C}'_u| - r + 1$ customers in $\mathcal{C}'_u$. Assume for a contradiction that there is no interval r-gathering where $\mathcal{C}''_u$ are assigned to f_u. Let A' be an interval r-gathering with maximum number of customers $\mathcal{D}_u \subset \mathcal{C}''_u$ assigned to f_u. Let $C_s \in \mathcal{C}''_u$ be the customer with smallest $t(C_s)$ which is not assigned to f_u. Let C_s is assigned to $f_v \geq f_u$. By Lemma 4, any customer $C_t \in \mathcal{C}''_u$ with $t(C_t) \geq t(C_s)$ is not assigned to f_u. We first claim that the number of customers assigned to f_v is exactly r. Otherwise we can reassign C_s to f_u and thus contradicting our assumption. Let $\mathcal{C}'_v$ be the customers assigned to f_v. We now claim that there is an interval r-gathering where $\mathcal{C}'_v$ consists of r customers having consecutive right end-points in $\mathcal{C}_u$. Assume otherwise for a contradiction. Let A'' be an interval r-gathering with minimum number of reverse pairs where a reverse pair is a pair of customer C_x, C_y with $t(C_x) \leq t(C_y)$, C_y assigned to f_v, C_x assigned to $f_w > f_v$. Since $t(C_x) \leq t(C_y)$ and $f_v \leq f_w$, we get $s(C_y) \leq f_w \leq t(C_y)$. We now derive a new interval r-gathering by reassigning C_x to f_v and C_y to f_w, a contradiction. Now since $|\mathcal{D}_u| < |\mathcal{C}'_u| - r + 1$, we get $|\mathcal{C}'_u \setminus \mathcal{D}_u| \geq r$. Thus $\mathcal{C}'_v \subset \mathcal{C}'_u$. We now derive a new interval r-gathering by assigning $\mathcal{C}'_v$ to f_u. A contradiction. □

We now give an algorithm **Interval-r-gather** for the interval r-gathering problem.

We now have the following theorem.

Theorem 1. *The algorithm* **Interval-r-gather** *decides whether there is an interval r-gathering of $\mathcal{C}$ to F, and constructs one if exists in $O(m + n\log n + nr^{\frac{n}{r}})$ time.*

Proof. The correctness of Algorithm Interval-r-gather is immediate from lemma 4 and 5.

We now estimate the running time of the algorithm. We can sort the customers based on their right end-points in $O(n \log n)$ time. For each customer we can precompute the preceding facility f_u in $O(n + m)$ time. For each facility f_u we can precompute the sets of customers C_u containing each facility and the leftmost customer C_j having left end-point on right of f_u in $O(n + m)$ time. In each call to Interval-r-gather, we need $O(|C_u|)$ time and at most r recursive calls to Interval-r-gather. Let $T(n)$ be the running time of the algorithm for n customers. We have $T(n) \leq O(|C_u|) + \sum_{i=1}^{r} T(n - r + 1) \leq O(nr^{\frac{n}{r}})$. Thus the running time of the algorithm is $O(m + n \log n + nr^{\frac{n}{r}})$. □

We now have the following theorem.

Theorem 2. *Let $\mathcal{C} = \{C_1, C_2, \cdots, C_n\}$ be a set of uncertain customers on a line each of which is specified by a piece-wise uniform function consisting of $k+1$ pieces, and $F = \{f_1, f_2, \cdots, f_m\}$ be a set of m facilities on the line. Then the optimal r-gathering can be constructed in $O(nk + mn\log n + (m + n\log k + n\log n + nr^{\frac{n}{r}})\log mn)$ time.*

Algorithm 1: Interval-r-gather($\mathcal{C}, F$)

```
Input  : A set C of customers each having an interval and a set F of facilities on a line
Output: An interval r-gathering if exists
if |C| < r or F = ∅ then
    return ∅;
endif
C_i ← leftmost customer in C;
f_u ← preceding facility of C;
C_u ← the set of customers containing f_u in their intervals;
C_j ← leftmost customer in C \ C_u;
C'_u ← the set of customers in C_u having smaller right end-point than t(C_j);
F' ← the set of facilities right to f;
if |C_u| < r then
    return ∅;
endif
if |C'_u| < r then
    D_u ← the set of r leftmost customers in C_u;/* Lemma 5(a) */
    A ← Assignment of D_u to f_u;
    Ans ← Interval-r-gather(C \ D_u, F');
    if Ans ≠ ∅ then
        return Ans ∪ A;
    endif
    return ∅;
endif
D_u ← the set of max{r, |C'_u| − r + 1} leftmost customers in C_u; /* Lemma 5(b) */
A ← Assignment of D_u to f_u;
C''_u ← C'_u \ D_u;
while C''_u is not empty do
    Ans ← Interval-r-gather(C \ D_u, F');
    if Ans ≠ ∅ then
        return Ans ∪ A;
    endif
    C_k ← leftmost customer in C''_u; /* (possibly with more customers) */
    A' ← Assignment of C_k to f_u;
    A ← A ∪ A';
    D_u ← D_u ∪ {C_k};
    C''_u ← C''_u \ {C_k};
end
return ∅ ;
```

Proof. We give outline of an algorithm to compute optimal r-gathering. We first compute the $E[d(p, C_i)]$ function for each $C_i \in \mathcal{C}$. This takes $O(nk)$ time in total. By Corollary 1, we compute the sorted list of all expected distances between customers and facilities in $O(nk+mn\log n)$ time. We find the optimal r-gathering by binary search, using the $O(m+n\log n+nr^{\frac{n}{r}})$ time algorithm for interval r-gathering $\log mn$ times. For each r-interval gathering problem, we compute the (C_i, b)-intervals in $O(n\log k)$ time. Thus finding optimal r-gathering by binary search requires $O(nk + mn\log n + (m + n\log k + n\log n + nr^{\frac{n}{r}})\log mn)$ time. □

3.2 Uniform Distribution

In this section we give an algorithm for the uncertain r-gathering problem when each customer location is specified by a well-separated uniform distribution.

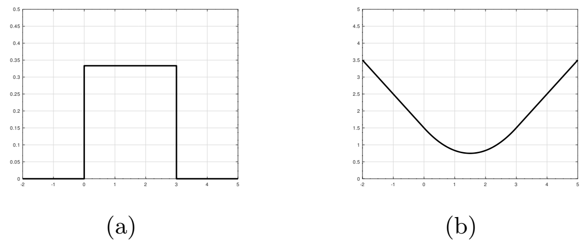


Fig. 2. (a) Illustration of a uniform distribution and (b) corresponding function of expected distance.

In the uniform distribution model, location of each customer C_i is specified by a function $g_i : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$ where $g_i(p) = 1/(t_i - s_i)$ if $s_i \le p \le t_i$ and $g_i(p) = 0$ otherwise. We denote the uniform distribution between $[s_i, t_i]$ by $U(s_i, t_i)$. The customer C_i having a uniform distribution $U(s_i, t_i)$ is denoted by $C_i \sim U(s_i, t_i)$. Figure 2(a) illustrates a uniform distribution where $s_i = 0$ and $t_i = 3$. The range of $U(s_i, t_i)$, denoted by l_i, is the value of $t_i - s_i$, and the mean of $U(s_i, t_i)$, denoted by μ_i, is the value of $\frac{s_i+t_i}{2}$. The uniform distribution model is a special case of the histogram model described in Section 3.1. We now have the following lemma.

Lemma 6. *Let $C \sim U(s,t)$ be an uncertain point. Then the function $E[d(p,C)]$ consists of a parabola in the interval $[s,t]$ and two straight lines of slope +1 and -1 in interval (t, ∞) and $(-\infty, s)$, respectively. Furthermore the minimum value of $E[d(p,C)]$ is $\frac{l}{4}$ and the value of $E[d(p,C)]$ at s, t is $\frac{l}{2}$.*

Proof. We use the Equation 1 to compute the function $E[d(p,C)]$.

$$E[d(p,C)] = \begin{cases} \mu - p & \text{if } p < s \\ \frac{1}{l}(p-\mu)^2 + \frac{l}{4} & \text{if } s \le p \le t \\ -\mu + p & \text{if } p > t \end{cases} \tag{2}$$

At $p = s$ we get $E[d(s,C)] = \frac{1}{t-s}\left(s - \frac{s+t}{2}\right)^2 + \frac{t-s}{4} = \frac{t-s}{2} = \frac{l}{2}$. Similarly, $E[d(t,C)] = \frac{l}{2}$. Now for $p < s$ and $p > t$, $E[d(p,C)] \ge \frac{t-s}{2}$. The minimum value of the parabola $\frac{1}{t-s}\left(p - \frac{s+t}{2}\right)^2 + \frac{t-s}{4}$ is $\frac{l}{4}$ at $p = \frac{s+t}{2}$. □

We have the following lemma.

Lemma 7. *Let $C \sim U(s,t)$ be an uncertain point and b be a number. Then the (C,b)-interval can be computed in $O(1)$ time.*

Proof. To find the (C,b)-interval, we first compute the inverse of the Equation 2. For $E[d(p,C)] = b > \frac{l}{2}$, we have $p < s$ or $p > t$. Thus we get, $p = \mu \pm b$. For $\frac{l}{4} \le E[d(p,C)] = b \le \frac{l}{2}$, we have $s \le p \le t$. Thus we get $p = \mu \pm \sqrt{l(b - \frac{l}{4})}$. Finally there is no p for which $E[d(p,C)] < \frac{l}{4}$. Hence the (C,b)-interval for $b < \frac{l}{4}$ is empty. Thus the (C,b)-interval I can be written as following.

$$I = \begin{cases} [\mu - b, \mu + b] & \text{if } b > \frac{l}{2} \\ [\mu - \sqrt{l(b-\frac{l}{4})}, \mu + \sqrt{l(b-\frac{l}{4})}] & \text{if } \frac{l}{4} \le b \le \frac{l}{2} \\ \emptyset & \text{if } b < \frac{l}{4} \end{cases} \quad (3)$$

By Equation 3 we can compute (C,b)-interval in $O(1)$ time. □

Let $C_i \sim U(s_i,t_i), C_j \sim U(s_j,t_j)$ be two uncertain points. Let $l_{max} = \max\{l_i,l_j\}$ and $l_{min} = \min\{l_i,l_j\}$. We call C_i, C_j *well-separated* if none of the intervals $[s_i,t_i]$ and $[s_j,t_j]$ is contained within the other and $|\mu_i - \mu_j| \ge \frac{1}{2}\sqrt{l_{min}(l_{max}-l_{min})}$.

Lemma 8. *Let $C_i \sim U(s_i,t_i), C_j \sim U(s_j,t_j)$ be two uncertain well-separated points and b be a number. Let I_i, I_j be the (C_i,b)-interval and (C_j,b)-interval respectively. Then none of I_i and I_j is contained in the other.*

Proof. Omitted. □

If the customer locations are specified by well-separated uniform distributions, we can solve the decision version of uncertain r-gathering problem by dynamic programming as follows. A subproblem asks to determine whether there is an r-gathering with cost at most b for the set of customers $C_1, C_2, \cdots, C_i$. Thus we have at most n distinct subproblems, and to solve a subproblem we need to check n smaller subproblems, so we can design an $O(m+n^2)$ time algorithm.

We can improve the running time as follows. A subproblem $P(i)$ asks to find a set of customers $\mathcal{C}_i$ and an interval r-gathering A of customers $\mathcal{C}_i \subseteq \mathcal{C}$ to $F_i = \{f_1, f_2, \cdots, f_i\}$ such that (1) $\mathcal{C}_i$ contains every customer C_i with $t(C_i) \le f_i$ (possibly with more customers), (2) f_i serves at least r customers, and (3) $\max_{C\in\mathcal{C}_i}\{t(C)\}$ is minimum. Let $C_{z(i)}$ be the customer with $\max_{C\in\mathcal{C}_i}\{t(C)\}$. We can observe that there is a proper interval r-gathering of $\mathcal{C}$ to F if and only if some $P(i)$ with $f_i \ge s(C_n)$ has a solution.

Lemma 9. *If $P(i)$ has a solution, then there is an interval r-gathering where customers assigned to each open facility have consecutive right end-points.*

Proof. Omitted. □

We now have the following lemma.

Lemma 10. *If $P(i)$ and $P(j)$ have solutions and $i < j$, then $t(C_{z(i)}) \le t(C_{z(j)})$.*

Proof. For a contradiction assume $t(C_{z(i)}) > t(C_{z(j)})$. Let A_j be an interval r-gathering corresponding to $P(j)$. Since all the intervals are proper, we have $s(C_{z(i)}) > s(C_{z(j)})$, and $s(C_{z(j)}) \le f_i$. Let $\mathcal{C}'_j$ be the set of customers assigned to any facility between f_i to f_j (including f_i, f_j) in A_j. For any customer $C_k \in \mathcal{C}'_j$, we have $s(C_k) \le f_i$ and $t(C_k) \ge f_i$. We now derive a new interval r-gathering A'_j by reassigning the leftmost r customers $\mathcal{C}'_j$ to f_i. Clearly, $\max_{C\in\mathcal{C}'_j}\{t(C)\} < t(C_{z(i)})$ and thus A'_j is a solution of $P(i)$, a contradiction. □

Using Lemma 9 and 10, we can determine whether $P(i)$ has solution or not. We have two cases. If $f_i \le t(C_1)$, then $P(i)$ may have a solution with exactly one open facility f_i, and the solution exists if and only if f_i is contained within at least r intervals. Otherwise $f_i > t(C_1)$, then $P(i)$ may have a solution with two or more open facilities. In this case $P(i)$ has a solution if and only if for some $j < i$ $P(j)$ has a solution, there is no customer C with $f_j < s(C) \le t(C) < f_i$, and there are at least r customers in $\mathcal{C} \setminus \mathcal{C}_j$ containing f_i. Intuitively f_j is a possible second rightmost open facility in a solution of $P(i)$.

We fix the $P(j)$ with minimum j, if $P(i)$ has a solution, and we say f_j the mate of f_i, and denoted as $mate(f_i)$. We have the following lemma.

Lemma 11. *If $P(i)$ and $P(i+1)$ have solutions, then $mate(f_i) \leq mate(f_{i+1})$.*

Proof. For a contradiction assume $mate(f_i) > mate(f_{i+1})$. Let $f_j = mate(f_i)$ and $f_{j'} = mate(f_{i+1})$. By Lemma 10 we have $t(C_{z(j)}) \geq t(C_{z(j')})$. Since $f_{j'}$ is mate of f_{i+1}, there is no customer C such that $f_{j'} < s(C) \leq t(C) < f_{i+1}$. If $t(C_{z(j)}) < f_i$, then $f_{j'}$ is also a mate of f_j, a contradiction. Now if $t(C_{z(j)}) \geq f_j$, then $f_{j'}$ is a mate of f_j since $t(C_{z(j')}) \leq t(C_{z(j)})$, a contradiction. □

We now have the following lemma.

Lemma 12. *Let f_i be a facility with $f_i > t(C_1)$ and for some $j < i$, $P(j)$ has a solution, and $\mathcal{C} \setminus \mathcal{C}_j$ contains no customer C with $f_j < s(C)$ and $t(C) < f_i$. Fix the $P(j)$ with minimum j. Then the following holds.*
(a) If $\mathcal{C} \setminus \mathcal{C}_j$ has less than r customers containing f_i, then no facility $f_{j'}$ with $f_{j'} \geq f_j$ is a mate of f_i, and $P(i)$ has no solution.
(b) If $P(i+1)$ has a solution, then $mate(f_{i+1}) \geq f_j$.

Proof. (a) By Lemma 10 for any facility $f_{j'} \geq f_j$, if $P(j')$ has a solution, then $t(C_{z(j')}) \geq t(C_{z(j)})$. Thus the number of customers in $\mathcal{C} \setminus \mathcal{C}_{j'}$ containing f_i in their interval is less than r.
(b) Assume for a contradiction that $mate(f_{i+1}) \leq f_j$. Let $f_{i'} = mate(f_{i+1})$. Thus there is no customer C with $f_{i'} < s(C)$ and $t(C) < f_{i+1}$. Since $f_{i'} \leq f_i \leq f_{i+1}$, there is no customer C such that $f_{i'} < s(C)$ and $t(C) < f_i$. Hence, $f_{i'}$ is the leftmost facility such that $P(i')$ has a solution and there is no customer C with $f_{i'} < s(C)$ and $t(C) < f_i$, a contradiction. □

By Lemma 11 and 12, we observe that we can search for $mate(f_{i+1})$ from where the search for mate of $mate(f_i)$ ends. We now give the following Algorithm called **Proper-interval-r-gather**.

If the intervals are sorted according to their right end-points and the facilities are ordered from left to right, then we can preprocess the set of customers containing each facility in linear time. Each customer and each facility have to be processed for a constant number of times. Hence the algorithm runs in $O(n+m)$ time. We thus have the following theorem.

Theorem 3. *Let $F = \{f_1, f_2, \cdots, f_m\}$ be a set of facilities on a line and $\mathcal{C} = \{C_1, C_2, \cdots, C_n\}$ be a set of customers where each customer C_i has an interval $I_i = [s(C_i), t(C_i)]$ and no interval is contained within any other interval. The algorithm* **Proper-interval-r-gather** *decides whether there is an interval r-gathering of $\mathcal{C}$ to F, and constructs one if exists in $O(n+m)$ time.*

We now give outline of the algorithm to solve uncertain r-gathering problem on a line where the customer locations are specified by well-separated uniform distributions. Computing the function $E[d(p, C_i)]$ for all the customers takes $O(n)$ time. We can compute the expected distances between customer C_i and all the facilities in $O(m)$ time. Since the function $E[d(p, C_i)]$ is unimodal, the expected distances between C_i and all the facilities can be sorted in $O(m)$ time. Computing the expected distances between each pair of customers and facilities takes $O(mn)$ time and we can merge the of n sorted list of expected distances in $O(mn \log n)$ time using heap. We do binary search on the ordered list of expected distances to find the optimal r-gathering. Given b we can compute the (C, b)-intervals for all customers in $O(n)$ time. The (C, b)-intervals can be sorted in $O(n \log n)$ time. Solving each decision instance takes $O(m+n)$ time. Thus to find the optimal solution by binary search we need to solve the decision instances $\log mn$ times, so $O((n \log n + m + n) \log mn)$ in total. Hence the running time is $O(mn \log n + (n \log n + m) \log mn)$. Thus we have the following theorem.

Theorem 4. *Let $F = \{f_1, f_2, \cdots, f_m\}$ be a set of facilities on a line and $\mathcal{C} = \{C_1, C_2, \cdots, C_n\}$ be a set of customers where each customer C_i has a well-separated uniform distribution. Then an optimal r-gathering of $\mathcal{C}$ to F can be constructed in $O(mn \log n + (n \log n + m) \log mn)$ time.*

Algorithm 2: Proper-interval-r-gather($\mathcal{C}, F$)

```
Input  : A set C of customers each having an interval where no interval is contained
         within other, a set of F of facilities on the line
Output: An interval r-gathering if exists
if |C| < r or F = ∅ then
  | return ∅;
endif
i ← 1;
/* One open facility */
while f_i ≤ t(C_1) do
  | if f_i ≥ s(C_r) then
  |   | z(i) ← r;
  | endif
  | i ← i + 1;
end
j ← 1;
/* Two or more open facilities */
while i ≤ m do
  | C_i ← {C_1, C_2, ··· , C_z(i)};
  | while j ≤ i do
  |   | if C \ C_j has at least r customers containing f_i and C \ C_j has no customer C with
  |   |   f_j < s(C) and t(C) < f_i then
  |   |   | z(i) ← index of the r-th customer in C \ C_j containing f_i; /* P(i) has a
  |   |   |   solution */
  |   |   | mate(i) ← j;
  |   |   | break;
  |   | endif
  |   | if There is no customer between f_j and f_i, and C \ C_j has less than r customers
  |   |   containing f_i then
  |   |   | break; /* P(i) has no solution, Lemma 12(a) */
  |   | endif
  |   | j ← j + 1;
  | end
  | i ← i + 1;
end
if Some P(i) with f_i ≥ s(C_n) has a solution then
  | Compute an interval r-gathering A of C to F;
  | return A;
endif
return ∅;
```

4 Conclusion

In this paper we presented an $O(nk + mn\log n + (m + n\log k + n\log n + nr^{\frac{n}{r}})\log mn)$ time algorithm for the one-dimensional uncertain r-gathering problem when the customers are given by piecewise uniform functions. We also gave an $O(mn\log n+(n\log n+m)\log mn)$ time algorithm when the customers are given by well-separated uniform distributions.

References

1. Agarwal, P.K., Cheng, S., Tao, Y., Yi, K.: Indexing uncertain data. In: Proceedings of the Twenty-Eigth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2009. pp. 137–146 (2009)
2. Agarwal, P.K., Efrat, A., Sankararaman, S., Zhang, W.: Nearest-neighbor searching under uncertainty I. Discrete & Computational Geometry 58(3), 705–745 (2017)
3. Agarwal, P.K., Har-Peled, S., Suri, S., Yildiz, H., Zhang, W.: Convex hulls under uncertainty. Algorithmica 79(2), 340–367 (2017)
4. Ahmed, S., Nakano, S., Rahman, M.S.: r-gatherings on a star. In: Proceedings of the 13th International Workshop on Algorithms and Computation. vol. 11355 of Lecture Notes in Computer Science, pp. 31–42. Springer Nature Switzerland (2019)
5. Akagi, T., Nakano, S.: On r-gatherings on the line. In: Proceedings of Frontiers in Algorithmics. vol. 9130 of Lecture Notes in Computer Science, pp. 25–32. Springer International Publishing, Cham (2015)
6. Armon, A.: On min-max r-gatherings. Theoretical Computer Science 412(7), 573 – 582 (2011)
7. Drezner, Z., Hamacher, H.W.: Facility Location: Applications and Theory. Springer, New York (2004)
8. Guha, S., Meyerson, A., Munagala, K.: Hierarchical placement and network design problems. In: Proceedings 41st Annual Symposium on Foundations of Computer Science. pp. 603–612 (2000)
9. Han, Y., Nakano, S.: On r-gatherings on the line. In: Proceedings of FCS 2016. pp. 99–104 (2016)
10. Kamousi, P., Chan, T.M., Suri, S.: Closest pair and the post office problem for stochastic points. Computational Geometry 47(2), 214–223 (2014)
11. Karget, D.R., Minkoff, M.: Building steiner trees with incomplete global knowledge. In: Proceedings 41st Annual Symposium on Foundations of Computer Science. pp. 613–623 (2000)
12. Nakano, S.: A simple algorithm for r-gatherings on the line. In: Proceedings of WALCOM: Algorithms and Computation. vol. 10755 of Lecture Notes in Computer Science, pp. 1–7. Springer International Publishing, Cham (2018)
13. Sarker, A., Sung, W., Rahman, M.S.: A linear time algorithm for the r-gathering problem on the line (extended abstract). In: Proceedings of WALCOM: Algorithms and Computation. vol. 11355 of Lecture Notes in Computer Science, pp. 56–66. Springer Nature Switzerland, Cham (2019)
14. Snyder, L.V.: Facility location under uncertainty: a review. IIE Transactions 38(7), 547–564 (2006)
15. Suri, S., Verbeek, K.: On the most likely voronoi diagram and nearest neighbor searching. International Journal of Computer Geometry Applications 26(3-4), 151–166 (2016)
16. Tao, Y., Xiao, X., Cheng, R.: Range search on multidimensional uncertain data. ACM Transaction on Database Systems 32(3), 15 (2007)
17. Wang, H., Zhang, J.: One-dimensional k-center on uncertain data. Theoretical Computer Science 602, 114–124 (2015)
18. Yiu, M.L., Mamoulis, N., Dai, X., Tao, Y., Vaitis, M.: Efficient evaluation of probabilistic advanced spatial queries on existentially uncertain data. IEEE Transaction on Knowledge and Data Engineering 21(1), 108–122 (2009)