SIAM J. ALG. DISC. METH.
Vol. 4, No. 2, June 1983

0196-5212/83/0402-0005 $01.25/0

FINDING LEAST-DISTANCES LINES*

NIMROD MEGIDDO† AND ARIE TAMIR‡

Abstract. We consider the following problem related to both location theory and statistical linear regression. Given n points in the plane find a straight line L so as to minimize the weighted sum of the distances of the points to L relative to either the Euclidean metric or the l_1-metric. We present $O(n^2 \log n)$ and $O(n \log^2 n)$ time algorithms for the Euclidean and rectilinear cases, respectively.

1. Introduction. We consider the following problem which is related to both location theory and statistics: Given n points in the plane $(x_1, y_1), \cdots, (x_n, y_n)$ together with positive weights $w_1, \cdots, w_n$, find a straight line L so as to minimize $\sum_{i=1}^{n} w_i d(x_i, y_i; L)$, where d is the distance function from L relative to either the Euclidean metric or the l_1-metric.

The location theory aspects of the problem are obvious. One may think of locating a portion of a new railroad so as to minimize the average cost to the users who have to reach the tracks from different small towns. The problem is also closely related to linear regression, with the difference that here we minimize the sum of distances instead of the squared distances. The latter case is computationally much easier since there are easy formulas available for the least-squares line. This is true both in the case where the distance is measured parallel to one of the axes and also when the distance is measured vertically to the line.

We note that the problem is related to the classic Weber problem [5], [13]. The Weber problem is to find a single point so as to minimize the average distance from it to n given points. When the problem is posed with respect to the Euclidean metric no polynomial time algorithms are known even when all the weights are equal. Relative to the l_1-metric the Weber problem is separable into two one-dimensional problems and hence is solvable in linear time by a weighted-median-finding algorithm [1].

Following the terminology of location theory we call our problem the 1-line median problem. We present in this paper an $O(n^2 \log n)$ algorithm for the Euclidean problem and an $O(n \log^2 n)$ algorithm for the rectilinear problem.

2. The Euclidean problem. In this section we focus on the Euclidean case. It is easy to see that a 1-line median can always be chosen so as to contain one of the n given points. This is because a parallel translation of the line which contains none of the points results in a linear change in the objective function as long as none of the points is reached. We, however, claim that a 1-line median can be chosen so as to contain at least two of our n points. This will enable us to consider only a set of $O(n^2)$ candidate lines for the 1-line median.

LEMMA 1. *Relative to the Euclidean metric there exists a 1-line median which contains at least two points from the set* $\{(x_1, y_1), \cdots, (x_n, y_n)\}$.

Proof. We have already argued that at least one point lies on the line. Thus, we assume without loss of generality that the point (x_1, y_1) lies on the line. Moreover, we may translate the coordinate system so that $x_1 = y_1 = 0$. In other words, we may pose our problem as of finding a straight line of the form $ax + by = 0$ which minimizes the sum of weighted distances from the points (x_i, y_i) $(i = 2, \cdots, n)$ to the line. The

* Received by the editors November 30, 1981, and in revised form July 14, 1982.

† Department of Statistics, Tel Aviv University, Tel Aviv 69978, Israel. This author's work was supported in part by the National Science Foundation under grant ECS8121741.

‡ Department of Statistics, Tel Aviv University, Tel Aviv 69978, Israel.

distance between a point (x_i, y_i) and a line $ax+by=0$ $(a^2+b^2\neq 0)$ is equal to $(a^2+b^2)^{-1/2}|ax_i+by_i|$ so, formally, we now wish to minimize the function $f(a,b)=\sum_{i=2}^{n} w_i|ax_i+by_i|$ subject to the constraint $a^2+b^2=1$.

Suppose (a^*, b^*) is an optimal solution for the optimization problem we have posed. Let $S^+=\{i: 2\leqq i\leqq n,\ a^*x_i+b^*y_i\geqq 0\}$ and $S^-=\{i: 2\leqq i\leqq n,\ a^*x_i+b^*y_i\leqq 0\}$. It follows that

$$f(a^*,b^*)=\sum_{i\in S^+} w_i(a^*x_i+b^*y_i)-\sum_{i\in S^-} w_i(a^*x_i+b^*y_i)$$

$$=\Big(\sum_{i\in S^+} w_ix_i-\sum_{i\in S^-} w_ix_i\Big)a^*+\Big(\sum_{i\in S^+} w_iy_i-\sum_{i\in S^-} w_iy_i\Big)b^*.$$

Let α and β denote the coefficients of a^* and b^*, respectively, in the latter equality, i.e., $f(a^*,b^*)=\alpha a^*+\beta b^*$. A necessary condition for (a^*, b^*) to minimize $f(a,b)$ (subject to $a^2+b^2=1$) is that it is also an optimal solution for the following optimization problem:

$$\begin{aligned}&\underset{a,b}{\text{minimize}}\ \alpha a+\beta b,\\ &\text{s.t. } ax_i+by_i\geqq 0 \qquad (i\in S^+),\\ &\qquad ax_i+by_i\leqq 0 \qquad (i\in S^-),\\ &\qquad a^2+b^2=1.\end{aligned}$$

If $a^*x_i+b^*y_i=0$ for some i $(2\leqq i\leqq n)$, then the lemma holds since the line $a^*x+b^*y=0$ passes through (x_1, y_1) and (x_i, y_i). Thus, assume $a^*x_i+b^*y_i\neq 0$ for all i $(i=2,\cdots,n)$. We now observe that the constraints $ax_i+by_i\geqq 0$ $(i\in S^+)$ and $ax_i+by_i\leqq 0$ $(i\in S^-)$ are not binding at the point (a^*, b^*). This implies that (a^*, b^*) is in fact an optimal solution for the problem of minimizing $\alpha a+\beta b$ subject only to $a^2+b^2=1$. We note that under the present assumptions $\alpha^2+\beta^2\neq 0$, since otherwise all the points are colinear, which in turn implies $a^*x_i+b^*y_i=0$ for all i. The latter optimization problem has a unique local minimum (a', b'), where $a'=-\alpha(\alpha^2+\beta^2)^{-1/2}$ and $b'=-\beta(\alpha^2+\beta^2)^{-1/2}$ and the corresponding objective-function value is $-(\alpha^2+\beta^2)^{1/2}$. Thus $(a^*, b^*)=(a', b')$ and hence $\alpha a^*+\beta b^*=-(\alpha^2+\beta^2)^{1/2}<0$. This however is a contradiction since $\alpha a^*+\beta b^*=\sum_{i=2}^{n} w_i|a^*x_i+b^*y_i|\geqq 0$. It follows that at least for one i $(2\leqq i\leqq n)$ $a^*x_i+b^*y_i=0$ and that completes the proof.

An obvious consequence of Lemma 1 is that a 1-line median can be found in $O(n^3)$ time: Enumerate all the $O(n^2)$ candidates and compute the weighted sum of distances in each case.

We now develop an $O(n^2\log n)$ algorithm for the 1-line median problem. The idea is to sort the candidate lines according to their slopes and then enumerate them in that order so that it takes only constant time to evaluate the sum of distances in each case. Let $-\infty<s_1\leqq s_2\leqq\cdots\leqq s_j\leqq\cdots\leqq s_m\leqq\infty$ denote these slopes and assume that together with each slope we have an associated pair of points.

A necessary condition for a line $ax+by+c=0$ to be a 1-line median is that it separates the set of points into two sets of approximately the same weight; more precisely, if $W=\sum_{i=1}^{n} w_i$, $T^+=\{i: ax_i+by_i>-c\}$ and $T^-=\{i: ax_i+by_i<-c\}$, then the necessary condition is that $\sum_{i\in T^+} w_i$, $\sum_{i\in T^-} w_i\leqq\frac{1}{2}W$. In other words, the number $-c$ has to be a weighted-median of the set $H=H(a,b)=\{ax_i+by_i\}$ of the "heights" of the different points above the line $ax+by=0$.

Obviously, for every pair (a, b) there is such a number c. Imagine that we increase the slope of our line continuously from $-\infty$ to $+\infty$, always selecting the number c so

as to satisfy the necessary condition. Consider the linear order induced on the set of points by heights relative to the line. This order changes only when the slope of the line coincides with one of the s_j's, in which case the ranks of the two points associated with the critical slope are interchanged. This observation enables us to keep track of the sets T^+, T^- as we continuously change the slope of the line. Specifically, the sets T^+, T^- change only when the pair of points involved in a critical slope consists of no more than one member from either set. We note that some of the critical slopes may coincide (if three or more points are colinear), however this does not affect the complexity of the algorithm since we traverse all the pairs of points in any case. Given a, b and the sets T^+, T^-, c may be redefined as $-\max\{ax_i+by_i: i\notin T^+\}$ and then the weighted sum of distances becomes

$$(a^2+b^2)^{-1/2}\left[\left(\sum_{i\in T^+} w_ix_i-\sum_{i\in T^-} w_ix_i\right)a+\left(\sum_{i\in T^+} w_iy_i-\sum_{i\in T^-} w_iy_i\right)b+\left(\sum_{i\in T^+} w_i-\sum_{i\in T^-} w_i\right)c\right].$$

Suppose that we keep track of the sets T^+ and T^- as well as the quantities

$$\sum_{i\in T^+} w_ix_i,\quad \sum_{i\in T^-} w_ix_i,\quad \sum_{i\in T^+} w_iy_i,\quad \sum_{i\in T^-} w_iy_i,\quad \sum_{i\in T^+} w_i,\quad \sum_{i\in T^-} w_i,\quad \max\{ax_i+by_i:_i\notin T^+\}$$

when we sweep the slopes in a nondecreasing order. Then it takes only $O(n^2)$ time to evaluate the objective function at all $O(n^2)$ critical slopes and choose the optimal slope. (To avoid the square-root operation we may instead maximize our objective function squared.)

3. The rectilinear problem. In the present section we consider the 1-line-median problem in the case where the distances are measured rectilinearly, i.e.,

$$d(x_i, y_i; x_j, y_j)=|x_i-x_j|+|y_i-y_j|.$$

It turns out that the distance between a line $\{ax+by+c=0\}$ and a point (x_i, y_i) is given simply by $|ax_i+by_i+c|/\max(|a|, |b|)$. In other words, if the slope of the line is between -1 and 1, then the distance is measured from the point to the line in parallel to the y-axis; otherwise it is measured in parallel to the x-axis. Thus, we can solve two problems: one in which all distances are measured in parallel to the y-axis and another one in which they are measured in parallel to the x-axis; we then select one of the two accordingly.

We shall now describe an algorithm for finding a straight line $y=ax+b$ so as to minimize $\sum_{i=1}^{k} w_i|y_i-ax_i-b|$. This problem resembles the problem of linear regression where we seek best fit in least squares. However, we do not have available a nice formula for this least-distances line like the one for the regression line. Nevertheless, the present case is more favorable than the Euclidean one due to convexity properties which are discussed below.

Let $f(a, b)=\sum_{i=1}^{n} w_i|y_i-ax_i-b|$ and $g(a)=\min_b f(a, b)$. Obviously, $f(a, b)$ is convex and this implies that $g(a)$ is convex.

We will find the minimum of $g(a)$. We note that the function $g(a)$ is linear on intervals between consecutive slopes of lines determined by two of the given points. Thus, $g(a)$ is piecewise linear with breakpoints only at these values. The latter can be proved along the lines of Lemma 1. It is easy to devise an $O(n^2\log n)$ algorithm like the one in § 2. We will, however, develop a more efficient algorithm, exploiting the convexity of g.

It is easy to verify that, given a, the number $b=b(a)$ which minimizes $f(a, b)$ is a weighted-median of the set $\{y_i-ax_i\}$. Thus, $g(a)$ can be evaluated in $O(n)$ time [1].

Furthermore, even if a is a breakpoint of g, we can evaluate the one-sided derivatives $g'_+(a)$, $g'_-(a)$ of g at a. This is carried out as follows. Let $S^- = \{i: y_i - ax_i < b\}$, $S^0 = \{i: y_i - ax_i = b\}$ and $S^+ = \{i: y_i - ax_i > b\}$. We know that $w(S^-)$, $w(S^+) \leqq \frac{1}{2}W$. (For any $X \subseteq \{1, 2, \cdots, n\}$, $w(X) = \sum_{i \in X} w_i$.) Consider the set S^0 with the order induced by the x_i's. According to our choice of b, $S^0 \neq \emptyset$. Thus, there exists an $i \in S^0$ such that the sets $S^{0-} = \{j \in S^0: x_j > x_i\}$ and $S^{0+} = \{j \in S^0: x_j < x_i\}$ satisfy $w(S^-) + w(S^{0-}) \leqq \frac{1}{2}W$ and $w(S^+) + w(S^{0+}) \leqq \frac{1}{2}W$. If $\varepsilon > 0$ is sufficiently small, then $b(a+\varepsilon) = y_i - (a+\varepsilon)x_i$. This implies that $g'_+(a) = \sum_{j \in S^- \cup S^{0-}} w_j x_j - \sum_{j \in S^+ \cup S^{0+}} w_j x_j$. S^{0+} and S^{0-} can be obtained in $O(n)$ time, [1], which is, therefore, also the time to compute $g'_+(a)$. The evaluation of the left-hand side derivative is analogous. Thus we conclude that for a given a, it takes $O(n)$ time to compute $g(a)$, $g'_+(a)$ and $g'_-(a)$.

Let a^* denote the slope of the 1-line-median. For any a if $g'_+(a) < 0$, then $a \leqq a^*$ and if $g'_-(a) > 0$, then $a \geqq a^*$; otherwise, $g'_-(a) \leqq 0 \leqq g'_+(a)$ and that implies $a = a^*$. This enables us to search for a^* efficiently.

We will search for a' by applying a general method for solving parametric combinatorial problems first introduced in [6]. Efficient implementations are achieved with the aid of parallel computation algorithms as explained in [7]. The application in the present case is as follows. We utilize a parallel sorting algorithm by Preparata [8] which employs $n \log n$ "processors" and runs in $O(\log n)$ time. We will sort the set $\{1, \cdots, n\}$ by the numbers $\{y_i - a^* x_i\}$ without actually knowing the value of a^*. Instead, throughout the process an interval $[\alpha, \beta]$ such that $\alpha \leqq a^* \leqq \beta$ will be maintained. At any stage, the interval will have the property that the outcomes of all the comparisons executed so far will be independent of a provided $a \in [\alpha, \beta]$. Finally, the entire order will be constant over the current interval.

Suppose that we sort the set $\{y_i - ax_i\}$, where a is restricted to some interval $[\alpha, \beta]$, but unspecified yet. Then, when we need to compare some $y_i - ax_i$ with $y_j - ax_j$, the ratio $a' = (y_i - y_j)/(x_i - x_j)$ becomes critical for that comparison. However, we can test in $O(n)$ time whether $a' \geqq a^*$ or $a' \leqq a^*$ and update the interval accordingly. Corresponding to each step in Preparata's sorting scheme, there will be $n \log n$ such critical values produced, one by each processor. We can search the set of critical values for a^*, namely, we will perform a binary search until our interval is narrowed down so that it does not contain any critical value in its interior. This binary search requires $O(\log n)$ tests, where each test decides whether a critical point is to the left or to the right of a^*. Thus a single stage requires $O(n \log n)$ time. However, the entire sort runs in $O(\log n)$ stages, so that our algorithm finds in $O(n \log^2 n)$ time an interval $[\alpha_0, \beta_0]$ such that $a^* \in [\alpha_0, \beta_0]$ and $g(a)$ is linear over $[\alpha_0, \beta_0]$. Finding a^* is now straightforward.

To conclude this section we contrast our $O(n \log^2 n)$ algorithm with the different solution approaches to the problem which have appeared in the statistics literature. The first approach was to apply infinite iterative processes to find the least weighted absolute deviation line. References [4], [10] represent this approach. It should be noted that some of these iterative procedures do not even guarantee convergence (e.g., [10]). The second approach (e.g., [2], [3], [12]) was to formulate and solve the problem as a linear programming problem. These methods (which are also applicable to the multidimensional case) are finite, but it is not at all clear whether their bounds are polynomial in the number of points. To our knowledge, the method in [9], [11] is the only one which has a polynomial bound. Using our notation, the method amounts to the evaluation of all the breakpoints of the piecewise linear function $g(a)$, which are between some arbitrary value $\bar{a}$ and a^* (the minimum of $g(a)$). In the worst-case all the breakpoints of $g(a)$ may have to be looked at. Since no method is known to

perform this task in $o(n^2)$ time, our $O(n \log^2 n)$ algorithm improves considerably over all existing methods.

REFERENCES

[1] M. BLUM, R. W. FLOYD, V. R. PRATT, R. L. RIVEST AND R. E. TARJAN, *Time bounds for selection*, J. Comput. System Sci., 7 (1972), pp. 448–461.

[2] A. CHARNES, W. W. COOPER AND R. O. FERGUSON, *Optimal estimation of executive compensation by linear regression*, Management Sci., 1 (1955), pp. 138–151.

[3] W. D. FISHER, *A note on curve fitting with minimum deviations by linear programming*, J. Amer. Statist. Assoc., 56 (1961), pp. 359–362.

[4] O. J. KARST, *Linear curve fitting using least deviations*, J. Amer. Statist. Assoc., 53 (1958), pp. 118–132.

[5] H. W. KUHN AND R. E. KUENNE, *An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics*, J. Regional Sci., 4 (1962), pp. 21–33.

[6] N. MEGIDDO, *Combinatorial optimization with rational objective functions*, Math. Oper. Res., 4 (1979), pp. 414–424.

[7] N. MEGIDDO, *Applying parallel computation algorithms in the design of serial algorithms*, Proc. 22nd Annual IEEE Symposium on Foundations of Computer Science, 1981, pp. 399–408.

[8] F. P. PREPARATA, *New parallel-sorting schemes*, IEEE Trans. Comp., C-27 (1978), pp. 669–673.

[9] M. R. RAO AND V. SRINIVASAN, *A note on Sharpe's algorithm for minimizing the sum of absolute deviations in a simple regression problem*, Management Sci., 19 (1972), pp. 222–225.

[10] E. J. SCHLOSSMACHER, *An iterative technique for absolute deviations curve fitting*, J. Amer. Statist. Assoc., 68 (1973), pp. 857–859.

[11] W. G. SHARPE, *Mean-absolute deviation characteristic lines for securities and portfolios*, Management Sci., 18 (1971), pp. B1–B13.

[12] H. M. WAGNER, *Linear programming techniques for regression analysis*, J. Amer. Statist. Assoc., 54 (1959), pp. 206–212.

[13] A. WEBER, *Über Den Standrot der Industrien*, Tubingen, 1909.