Mathematical Programming 4 (1973) 98–107. North-Holland Publishing Company

A NOTE ON FERMAT'S PROBLEM

Harold W. KUHN *

Princeton University, Princeton, N.J., U.S.A.

Received 17 July 1972

The General Fermat Problem asks for the minimum of the weighted sum of distances from m points in n-space. Dozens of papers have been written on variants of this problem and most of them have merely reproduced known results. This note calls attention to the work of Weiszfeld in 1937, who may have been the first to propose an iterative algorithm. Although the same algorithm has been rediscovered at least three times, there seems to be no completely correct treatment of its properties in the literature. Such a treatment, including a proof of convergence, is the sole object of this note. Other aspects of the problem are given scant attention.

1. Introduction

The following optimization problem has fascinated mathematicians for over 300 years since it was first proposed by Fermat early in the 17th century: Given three points in the plane, find a fourth point such that the sum of the distances to the three given points is a minimum. The problem was generalized by Simpson in his *Doctrine and Application of Fluxions* (London, 1750) to asking for the minimum weighted sum of distances from three given points. In this note, we shall consider the same problem for m points in the Euclidean space E^n. Formally, let there be given m points $A_i = (a_{i1}, ..., a_{in})$, called *vertices*, and m positive numbers w_i, called *weights*. Furthermore, for $P = (x_1, ..., x_n)$, let

$$d_i(P) = \sqrt{\Sigma_j (x_j - a_{ij})^2}\ ,$$

the Euclidean distance from P to A_i, for $i = 1, ..., m$.

* This paper was written while the author was Science Faculty Fellow of the National Science Foundation at the London School of Economics.

1.1. *General Fermat Problem: Find a point that minimizes* $f(P) = \Sigma_i w_i\, d_i(P)$.

Although dozens of papers have been written on variants of this problem and most of them have merely reproduced known results, it seems that something new can be said about it. It is the purpose of this note to call attention to a little known work [5], which appears to have been the first to propose an iterative algorithm. Although the same algorithm has been rediscovered several times (see, for example, [1], [3] or [4]), there seems to be no completely correct treatment of its properties in the literature. Such a treatment, including a proof of convergence, is the sole object of this note; consequently, we shall give scant attention to other aspects of the problem. A more complete history of the problem and a statement and treatment of its dual are given in [2].

2. The algorithm

If the vertices A_i are not collinear, then f is positive and strictly convex in E^n. Hence the minimum of f is achieved at a *unique* point M. We shall only consider non-collinear problems; those problems excluded by this restriction are clearly trivial.

Motives both mathematical and physical (deriving from a string and weight model of the problem introduced by G. Pick as early as 1909) suggest the introduction of the negative of the gradient of f (i.e., the *resultant* of the forces in the strings). To this end, let

$$R(P) = \sum_i \frac{w_i}{d_i(P)} (A_i - P) \quad \text{if } P \neq A_i \text{ for all } i\,.$$

Obviously, R is not defined at any vertex A_i. However, by physical analogy, set

$$R_k = \sum_{i \neq k} \frac{w_i}{d_i(A_k)} (A_i - A_k) \quad \text{for } k = 1, ..., m\,,$$

and extend the definition of R by setting

$$R(A_k) = \max\{|R_k| - w_k, 0\}(R_k/|R_k|) \quad \text{for } k = 1, ..., m\,.$$

(Here, as elsewhere in this note, $|A|$ denotes the length of the vector A.) In the expression for $R(A_k)$, the length of R_k is compared with w_k. If $w_k \geqq |R_k|$, then $R(A_k) = 0$; otherwise, a "resultant" of magnitude $|R_k| - w_k$ is defined in the direction of R_k.

2.1. *The point $P = M$ if and only if $R(P) = 0$.*

Proof. If P is not a vertex, then the convexity and differentiability of f implies that the first-order conditions $R(P) = 0$ are both necessary and sufficient for a minimum.

If $P = A_k$, then consider a change from A_k to $A_k + tZ$ for $|Z| = 1$. Then direct calculation yields

$$\frac{\mathrm{d}}{\mathrm{d}t} f(A_k + tZ) = w_k - R_k \cdot Z \quad \text{for } t = 0 \text{ ,}$$

and hence the direction of greatest decrease of f from A_k is $Z = R_k / |R_k|$. (Here $A \cdot B$ denotes the inner product of A and B, and A^2 will be used as an abbreviation for $A \cdot A$.) Clearly, A_k is a local minimum if and only if

$$w_k - R_k^2 / |R_k| \geqq 0 \text{ ,}$$

which is the same as $R(A_k) = 0$. Again, the convexity of f implies that $R(A_k) = 0$ is both necessary and sufficient for A_k to be a global minimum.

2.2. *The point M is in the convex hull of the vertices A_i.*

Proof. If M is a vertex, then it is trivially in the convex hull. Otherwise, the condition $R(M) = 0$ yields the consequence

$$M = \sum_i \frac{w_i}{d_i(M)} A_i \Big/ \sum_i \frac{w_i}{d_i(M)} \text{ .}$$

Thus M is a weighted sum of the vertices with positive weights that sum to one.

The equation used in the proof of 2.2 suggests quite naturally a method of successive approximation. For $P \neq A_i$, $i = 1, ..., m$, define

$$T\colon P \to T(P) = \sum_i \frac{w_i}{d_i(P)} A_i \Big/ \sum_i \frac{w_i}{d_i(P)} .$$

For the sake of continuity, set $T(A_i) = A_i$ for $i = 1, ..., m$. We then have as an immediate corollary to 2.2:

2.3. *If $P = M$, then $T(P) = P$. If P is not a vertex and $T(P) = P$, then $P = M$.*

In effect, the algorithm proposed is merely a simple attempt to solve the first-order conditions $R(P) = 0$ iteratively. It seems to have been discovered in 1937 by Weiszfeld [5], who asserted that, for any P_0 that is not a vertex, the sequence $P_r = T^r(P_0)$ converges to M. In the next section, we shall investigate the properties of T and prove a corrected statement of this theorem.

3. Statement and proof of convergence

First, note that the algorithm proposed is a "long-step" gradient method. Indeed, recalling that $-R(P)$ is the gradient of f whenever it exists, direct calculation yields

$$T(P) = P + h(P)\, R(P) ,$$

where

$$h(P) = \prod_i d_i(P) \Big/ \sum_k (w_k \prod_{i \neq k} d_i(P))$$

for *all* points P. Thus the algorithm follows the direction of the resultant with precalculated length of step $h(P)\ |R(P)|$. Apart from the vertices, which are all left fixed by T, one difficulty with such methods is that they may "overshoot". The following result (first proved in [5]) shows that this is not the case.

3.1. *If $T(P) \neq P$, then $f(T(P)) < f(P)$.*

Proof. Since $T(P) \neq P$, P is not a vertex and

$$T(P) = \sum_i \frac{w_i}{d_i(P)} A_i \Big/ \sum_i \frac{w_i}{d_i(P)} \,.$$

This says that $T(P)$ is the center of gravity of weights $w_i/d_i(P)$ placed at the vertices A_i. Hence, by elementary calculus, $T(P)$ is the unique minimum of the strictly convex function

$$g(Q) = \sum_i \frac{w_i}{d_i(P)} d_i^2(Q) \,.$$

Since $P \neq T(P)$,

$$g(T(P)) = \sum_i \frac{w_i}{d_i(P)} d_i^2(T(P)) < g(P) = \sum_i \frac{w_i}{d_i(P)} d_i^2(P) = f(P) \,.$$

On the other hand,

$$g(T(P)) = \sum_i \frac{w_i}{d_i(P)} [d_i(P) + (d_i(T(P)) - d_i(P))]^2$$

$$= f(P) + 2(f(T(P)) - f(P)) + \sum_i \frac{w_i}{d_i(P)} [d_i(T(P)) - d_i(P)]^2 \,.$$

Combining these results,

$$2f(T(P)) + \sum_i \frac{w_i}{d_i(P)} [d_i(T(P)) - d_i(P)]^2 < 2f(P)$$

and the assertion $f(T(P)) < f(P)$ is proved.

A second possible difficulty with the algorithm is that the sequence of approximations might remain in the neighborhood of a non-optimal vertex. The following result shows that this cannot happen. Informally, it says that there is a neighborhood of each non-optimal vertex such that, if the approximation sequence enters it, then it is eventually "kicked out" by T.

3.2. *Suppose $A_k \neq M$. Then there exists $\delta > 0$ such that $0 < d_k(P) \leqq \delta$ implies $d_k(T^s(P)) > \delta$ and $d_k(T^{s-1}(P)) \leqq \delta$ for some positive integer s.*

Proof.

$$\begin{aligned} T(P) - A_k &= P + h(P)\,R(P) - A_k \\ &= h(P) \sum_{i \neq k} \frac{w_i}{d_i(P)} (A_i - P) + \left(\frac{h(P)\,w_k}{d_k(P)} - 1\right) (A_k - P). \end{aligned}$$

Since $A_k \neq M$, we have

$$\left| \sum_{i \neq k} \frac{w_i}{d_i(A_k)} (A_i - A_k) \right| > w_k \,.$$

Hence there exist $\delta' > 0$ and $\epsilon > 0$ such that

$$\left| \sum_{i \neq k} \frac{w_i}{d_i(P)} (A_i - P) \right| \geqq (1 + 2\epsilon)\, w_k \quad \text{for } d_k(P) \leqq \delta' \,.$$

By the definition of h, we have

$$\lim_{P \to A_k} \quad h(P)\, w_k / d_k(P) = 1 \,.$$

Hence there exists $\delta'' > 0$ such that

$$\left| \frac{h(P)\, w_k}{d_k(P)} - 1 \right| < \frac{\epsilon}{2(1 + \epsilon)} \quad \text{for } 0 < d_k(P) \leqq \delta'' \,.$$

Set $\delta = \min(\delta', \delta'')$. For $0 < d_k(P) \leqq \delta$, we have

$$\begin{aligned} d_k(T(P)) &> h(P)\,(1 + 2\epsilon)\, w_k - \frac{\epsilon}{2(1 + \epsilon)} d_k(P) \\ &> \left(1 - \frac{\epsilon}{2(1 + \epsilon)}\right)(1 + 2\epsilon)\, d_k(P) - \frac{\epsilon}{2(1 + \epsilon)} d_k(P) \\ &= (1 + \epsilon)\, d_k(P) \,. \end{aligned}$$

Since $d_k(P) > 0, (1+\epsilon)^t\, d_k(P) > \delta$ for some positive integer t and hence $d_k(T^s(P)) > \delta$ for some positive integer s with $d_k(T^{s-1}(P)) \leqq \delta$.

The following result (first proved in [5]), which could be used to derive 3.2, describes the behavior of T near all vertices, optimal or not.

3.3. $\lim_{P\to A_k} \{d_k(T(P))/d_k(P)\} = |R_k|/w_k$ *for* $k = 1, \ldots, m$.

Proof. For P not a vertex,

$$T(P) = \sum_i \frac{w_i}{d_i(P)} A_i \Big/ \sum_i \frac{w_i}{d_i(P)}$$

$$= \left(\sum_{i\neq k} \frac{w_i}{d_i(P)} (A_i - A_k) + A_k \sum_i \frac{w_i}{d_i(P)} \right) \Big/ \sum_i \frac{w_i}{d_i(P)} .$$

Hence

$$T(P) - A_k = \sum_{i\neq k} \frac{w_i}{d_i(P)} (A_i - A_k) \Big/ \sum_i \frac{w_i}{d_i(P)},$$

$$\frac{1}{d_k(P)} (T(P) - A_k) = \sum_{i\neq k} \frac{w_i}{d_i(P)} (A_i - A_k) \Big/ w_k \left(1 + \frac{d_k(P)}{w_k} \sum_{i\neq k} \frac{w_i}{d_i(P)} \right).$$

Taking the limits of the lengths of both sides,

$$\lim_{P\to A_k} \frac{d_k(T(P))}{d_k(P)} = \frac{|R_k|}{w_k} .$$

3.4. *Convergence Theorem: Given any* P_0, *define* $P_r = T^r(P_0)$ *for* $r = 1, 2, \ldots$. *If no* P_r *is a vertex, then* $\lim_{r\to\infty} P_r = M$.

Proof. With the possible exception of P_0, the sequence P_r lies in the convex hull of the vertices, a compact set. Hence, by the Bolzano–Weierstrasz Theorem, there exists at least one point P and a subsequence P_{r_l} such that $\lim_{l\to\infty} P_{r_l} = P$. To prove the theorem, we must verify that $P = M$ in all cases.

If $P_{r+1} = T(P_r) = P_r$ for some r, then the sequence repeats from that point and $P = P_r$. Since P_r is not a vertex, $P = M$ by 2.3.

Otherwise, by 3.1,

$$f(P_0) > f(P_1) > \dots > f(P_r) > \dots > f(M) .$$

Hence

$$\lim_{r\to\infty} (f(P_{r_l}) - f(T(P_{r_l})) = 0 .$$

Since the continuity of T implies

$$\lim_{l\to\infty} T(P_{r_l}) = T(P) ,$$

we have

$$f(P) - f(T(P)) = 0 .$$

Therefore, by 3.1, $P = T(P)$. If P is not a vertex, then $P = M$ by 2.3. In any event, P lies in the finite set of isolated points $\{A_1, \dots, A_m; M\}$, where M may be a vertex.

The only case that remains is $P = A_k$ for some k. If $A_k \neq M$, we first isolate A_k from the other vertices (and M if it is not a vertex) by a δ-neighborhood that satisfies 3.2. Then it is clear that we can choose our subsequence $P_{r_l} \to A_k$ such that $d_k(T(P_{r_l})) > \delta$ for all l. This means that the ratio $d_k(T(P_{r_l}))/d_k(P_{r_l})$ is unbounded. However, this contradicts 3.3. Hence $A_k = M$ and the theorem is proved.

The error in Weiszfeld's statement [5, p. 356] consists in ignoring the possibility that even if P_0 is chosen distinct from all vertices, some $P_r = T^r(P_0)$ may be a vertex. This may invalidate his arguments [5, pp. 362–363] where several quotients are then undefined. The following example shows that this is a real possibility and is a counterexample to Weiszfeld's theorem.

3.5. *Counterexample*: Consider the six vertices in the plane: $A_1 = (-2,0)$, $A_2 = (-1,0)$, $A_3 = (1,0)$, $A_4 = (2,0)$, $A_5 = (0,1)$, $A_6 = (0,-1)$, all with weights $w_i = 1$. The vertices are graphed in Fig. 1. Since the resultant vanishes at the origin, $M = (0,0)$. Consider the behavior of T on the segment from the origin to A_4. From the definition of R, it follows that $T((x_1,0)) = (x_1',0)$ for $0 \leqq x_1 \leqq 2$ and an elementary estimate

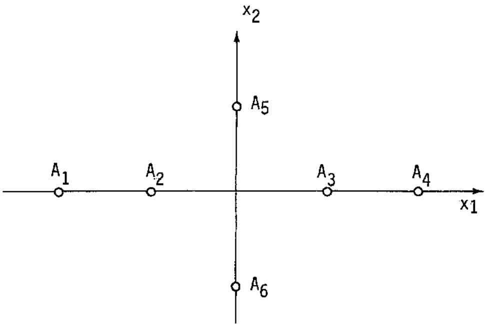

Fig. 1.

shows that $x_1' \leqq x_1$ on this interval. The behavior of T near A_3 and A_4 is provided by 3.3; the behavior near M can be established by an elementary calculation. The resulting graph is shown in Fig. 2. The important fact about this figure is that there is an x_0 (approximately 1.62) such that, for $P_0 = (x_0, 0)$ we have $T(P_0) = A_3$, which is not optimal. Thus, if one has the bad luck to start the algorithm from P_0, then $P_1 = A_3$ and the sequence repeats from that point. Thus the example shows that the sequence P_r need not converge to M.

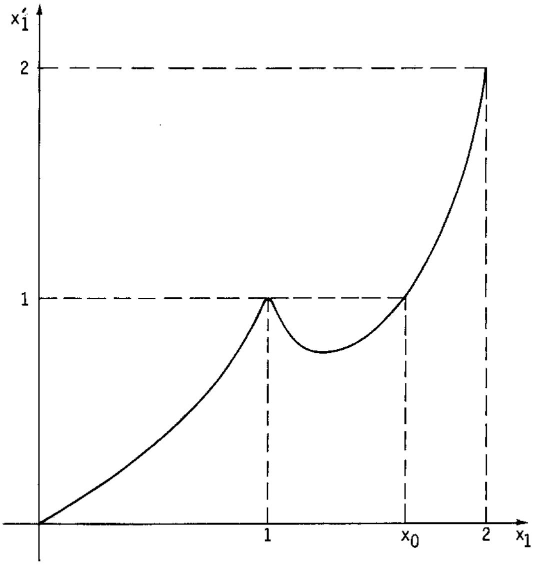

Fig. 2.

Of course, this is a very unlikely event. The following result expresses this precisely.

3.6. *For all but a denumerable number of P_0, $P_r = T^r(P_0)$ converges to M.*

Proof. The Convergence Theorem 3.4 establishes that, if no P_r is a vertex, then P_r converges to M. If we insert T from a vertex A_i, we must solve algebraic equations. Thus we obtain a finite number of P_0 such that $T(P_0) = A_i$. Hence, for a fixed positive r,

$$\{P_0 : T^r(P_0) = A_i \text{ for some } i = 1, ..., m\}$$

is finite. Finally,

$$\{P_0 : T^r(P_0) = A_i \text{ for some } i \text{ and } r\}$$

is denumerable.

References

[1] L. Cooper, "Location-allocation problems," *Operations Research* 11 (1963) 331–343.

[2] H.W. Kuhn, "On a pair of dual nonlinear programs," in: *Methods of nonlinear programming*, Ed. J. Abadie (North-Holland, Amsterdam, 1967) 38–54.

[3] H.W. Kuhn and R.E. Kuenne, "An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics," *Journal of Regional Science* 4 (1962) 21–33.

[4] W. Miehle, "Link-length minimization in networks," *Operations Research* 6 (1958) 232–243.

[5] E. Weiszfeld, "Sur le point pour lequel la somme des distances de *n* points donnés est minimum," *Tôhoku Mathematics Journal* 43 (1937) 355–386.