UCLA
UCLA Previously Published Works

Title
A Look at the Generalized Heron Problem through the Lens of Majorization-Minimization

Permalink

Journal
American Mathematical Monthly, 121(2)

ISSN
0002-9890

Authors
Chi, Eric C
Lange, Kenneth

Publication Date
2014

DOI
10.4169/amer.math.monthly.121.02.095

Peer reviewed

A Look at the Generalized Heron Problem through the Lens of Majorization-Minimization

Eric C. Chi and
Department of Human Genetics, University of California, Los Angeles, CA 90095,
ecchi@ucla.edu

Kenneth Lange
Departments of Human Genetics, Biomathematics, and Statistics, University of California, Los Angeles, CA 90095, klange@ucla.edu

Abstract

In a recent issue of this journal, Mordukhovich, Nam, and Salinas pose and solve an interesting non-differentiable generalization of the Heron problem in the framework of modern convex analysis. In the generalized Heron problem, one is given $k + 1$ closed convex sets in $\mathbb{R}^d$ equipped with its Euclidean norm and asked to find the point in the last set such that the sum of the distances to the first k sets is minimal. In later work, the authors generalize the Heron problem even further, relax its convexity assumptions, study its theoretical properties, and pursue subgradient algorithms for solving the convex case. Here, we revisit the original problem solely from the numerical perspective. By exploiting the majorization-minimization (MM) principle of computational statistics and rudimentary techniques from differential calculus, we are able to construct a very fast algorithm for solving the Euclidean version of the generalized Heron problem.

1 Introduction

In a recent article in this journal, Mordukhovich et al. [22] presented the following generalization of the classical Heron problem. Given a collection of closed convex sets $\{C_1, \ldots, C_k\}$ in $\mathbb{R}^d$, find a point $\mathbf{x}$ in the closed convex set $S \subset \mathbb{R}^d$ such that the sum of the Euclidean distances from $\mathbf{x}$ to C_1 through C_k is minimal. In other words,

$$\text{minimize } D(\mathbf{x}) := \sum_{i=1}^{k} d(\mathbf{x}, C_i) \text{ subject to } \mathbf{x} \in S, \quad (1)$$

where $d(\mathbf{x}, \Omega) = \inf\{\|\mathbf{x} - \mathbf{y}\| : \mathbf{y} \in \Omega\}$.

A rich history of special cases motivates this problem formulation. When $k = 2$, C_1 and C_2 are singletons, and S is a line, we recover the problem originally posed by the ancient mathematician Heron of Alexandria. The special case where $k = 3$; C_1, C_2, and C_3 are singletons; and $S = \mathbb{R}^2$ was suggested by Fermat nearly 400 years ago and solved by

Torricelli [13]. In his *Doctrine and Application of Fluxions*, Simpson generalized the distances to weighted distances. In the 19th century, Steiner made several fundamental contributions, and his name is sometimes attached to the problem [9, 11]. At the turn of the 20th century, the German economist Weber generalized Fermat's problem to an arbitrary number of singleton sets C_i. Weiszfeld published the first iterative algorithm[1] for solving the Fermat-Weber problem in 1937 [28, 29]. In the modern era, the Fermat-Weber problem has enjoyed a renaissance in various computational guises. Both the problem and associated algorithms serve as the starting point for many advanced models in location theory [18, 30].

The connections between celebrated problems such as the Fermat-Weber problem and the generalized Heron problem were noted earlier by Mordukhovich et al. [23]. In subsequent papers [21, 23], they generalize the Heron problem further to arbitrary closed sets, C_1, …, C_k and S in a Banach space. Readers are referred to their papers for a clear treatment of how one solves these abstract versions of the generalized Heron problem with state-of-the-art tools from variational analysis.

Here we restrict our attention to the special case of Euclidean distances presented by Mordukhovich et al. [23]. Our purpose is take a second look at this simple yet, in our opinion, most pertinent version of the problem from the perspective of algorithm design. Mordukhovich et al. [21, 22, 23] present an iterative subgradient algorithm for numerically solving problem (1) and its generalizations, a robust choice when one desires to assume nothing beyond the convexity of the objective function. Indeed, the subgradient algorithm works if the Euclidean norm is exchanged for an arbitrary norm. However, it is natural to wonder if there might be better alternatives for the finite-dimensional version of the problem with Euclidean distances. Here we present one that generalizes Weiszfeld's algorithm by invoking the majorization-minimization (MM) principle from computational statistics. Although the new algorithm displays the same kind of singularities that plagued Weiszfeld's algorithm [15], the dilemmas can be resolved by slightly perturbing problem (1), which we refer to as the generalized Heron problem for the remainder of this article. In the limit, one recovers the solution to the unperturbed problem. As might be expected, it pays to exploit special structure in a problem. The new MM algorithm is vastly superior to the subgradient algorithms in computational speed for Euclidean distances.

Solving a perturbed version of the problem by the MM principle yields extra dividends as well. The convergence of MM algorithms on smooth problems is well understood theoretically. This fact enables us to show that solutions to the original problem can be characterized without appealing to the full machinery of convex analysis dealing with non-differentiable functions and their subgradients. Although this body of mathematical knowledge is definitely worth learning, it is remarkable how much progress can be made with simple tools. The good news is that we demonstrate that crafting an iterative numerical solver for problem (1) is well within the scope of classical differential calculus. Our resolution can be understood by undergraduate mathematics majors.

---

[1]Kuhn [15] points out that Weiszfeld's algorithm has been rediscovered several times.

As a brief summary of things to come, we begin by recalling background material on the MM principle and convex analysis of differentiable functions. This is followed with a derivation of the MM algorithm for problem (1) and consideration of a few relevant numerical examples. We end by proving convergence of the algorithm and characterizing solution points.

2 The MM Principle

Although first articulated by the numerical analysts Ortega and Rheinboldt [24], the MM principle currently enjoys its greatest vogue in computational statistics [1, 17]. The basic idea is to convert a hard optimization problem (for example, non-differentiable) into a sequence of simpler ones (for example, smooth). The MM principle requires majorizing the objective function $f(\mathbf{y})$ by a surrogate function $g(\mathbf{y} \mid \mathbf{x})$ anchored at the current point $\mathbf{x}$. Majorization is a combination of the tangency condition $g(\mathbf{x} \mid \mathbf{x}) = f(\mathbf{x})$ and the domination condition $g(\mathbf{y} \mid \mathbf{x}) \quad f(\mathbf{y})$ for all $\mathbf{y} \in \mathbb{R}^d$. The associated MM algorithm is defined by the iterates

$$\mathbf{x}_{k+1} := \operatorname*{argmin}_{\mathbf{y} \in s} g(\mathbf{y} \mid \mathbf{x}_k). \quad (2)$$

Because

$$f(\mathbf{x}_{k+1}) \leq g(\mathbf{x}_{k+1} \mid \mathbf{x}_k) \leq g(\mathbf{x}_k \mid \mathbf{x}_k) = f(\mathbf{x}_k), \quad (3)$$

the MM iterates generate a descent algorithm driving the objective function downhill. Constraint satisfaction is enforced in finding $\mathbf{x}_{k+1}$. Under appropriate regularity conditions, an MM algorithm is guaranteed to converge to a local minimum of the original problem [16].

3 Background on Convex Analysis

As a prelude to deriving an MM algorithm, we review some basic facts from convex analysis in the limited context of differentiable functions. Deeper treatments can be found in the references [3, 4, 12, 25, 26]. Throughout this article, we denote the standard dot product between vectors $\mathbf{a}$ and $\mathbf{b}$ by $\langle \mathbf{a}, \mathbf{b} \rangle$. Recall that a differentiable function $f(\mathbf{y})$ is convex if and only if its domain S is convex and

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \quad (4)$$

for all $\mathbf{x}, \mathbf{y} \in S$. Provided $f(\mathbf{x})$ is twice differentiable, it is convex when its second differential $d^2 f(\mathbf{x})$ is positive semidefinite for all $\mathbf{x}$ and strictly convex when $d^2 f(\mathbf{x})$ is positive definite for all $\mathbf{x}$. These characterizations are a direct consequence of executing a second-order Taylor expansion of $f(\mathbf{y})$ and applying the supporting hyperplane inequality (4). The supporting hyperplane inequality (4) also leads to a succinct necessary and sufficient condition for a global minimum. A point $\mathbf{x} \in S$ is a global minimizer of $f(\mathbf{y})$ on S if and only if

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq 0 \quad (5)$$

for all $\mathbf{y} \in S$. Intuitively speaking, every direction pointing into S must lead uphill.

We conclude this section by reviewing projection operators [16]. Denote the projection of $\mathbf{x}$ onto a set $\Omega \subset \mathbb{R}^d$ by $P_\Omega(\mathbf{x})$. By definition, $P_\Omega(\mathbf{x})$ satisfies

$$P_\Omega(\mathbf{x}) := \arg\min_{\mathbf{y} \in \Omega} \| \mathbf{x} - \mathbf{y} \| .$$

If Ω is a closed convex set in $\mathbb{R}^d$, then $P_\Omega(\mathbf{x})$ exists and is unique. Furthermore, the projection operator is non-expansive in the sense that

$$\| P_\Omega(\mathbf{x}) - P_\Omega(\mathbf{y}) \| \leq \| \mathbf{x} - \mathbf{y} \|$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Non-expansion clearly entails continuity. Explicit formulas for the projection operator $P_\Omega(\mathbf{x})$ exist when Ω is a box, Euclidean ball, hyperplane, or halfspace. Fast algorithms for computing $P_\Omega(\mathbf{x})$ exist for the unit simplex, the ℓ_1 ball, and the cone of positive semidefinite matrices [10, 20].

The projection operator and the distance function are intimately related through the gradient identity $\nabla d(\mathbf{x}, C)^2 = 2[\mathbf{x} - P_C(\mathbf{x})]$. A standard proof of this fact can be found in reference [12, p. 181]. If $d(\mathbf{x}, C)^2 > 0$, then the chain rule gives

$$\nabla d(\mathbf{x}, C) = \nabla \sqrt{d(\mathbf{x}, C)^2} = \frac{\mathbf{x} - P_C(\mathbf{x})}{d(\mathbf{x}, C)}.$$

On the interior of C, it is obvious that $\nabla d(\mathbf{x}, C) = \mathbf{0}$. In contrast, differentiability of $d(\mathbf{x}, C)$ at boundary points of C is not guaranteed.

4 An MM Algorithm for the Heron Problem

Since it adds little additional overhead, we recast problem (1) in the Simpson form

$$\text{minimize } D(\mathbf{x}) := \sum_{i=1}^{k} \gamma_i d(\mathbf{x}, C_i) \text{ subject to } \mathbf{x} \in S \quad (6)$$

involving a convex combination of the distances $d(\mathbf{x}, C_i)$ with positive weights γ_i as suggested in [23]. We first derive an MM algorithm for solving problem (6) when $S \cap C_i = \emptyset$ for all i. This exercise will set the stage for attacking the more general case where S intersects one or more of the C_i. In practice, quadratic majorization is desirable because it promotes exact solution of the minimization step of the MM algorithm. It takes two successive majorizations to achieve quadratic majorization in our setting. The first is the simple majorization

$$d(\mathbf{x}, C_i) \leq \parallel \mathbf{x} - P_{C_i}(\mathbf{x}_m) \parallel$$

flowing directly from the definition of the distance function. The second is the majorization

$$\sqrt{u} \leq \sqrt{u_m} + \frac{1}{2\sqrt{u_m}}(u - u_m), \quad (7)$$

of the concave function $\sqrt{u}$ on the interval $(0, \infty)$. The combination of these two majorizations yields the quadratic majorization

$$d(\mathbf{x}, C_i) \leq \parallel \mathbf{x}_m - P_{C_i}(\mathbf{x}_m) \parallel + \frac{\parallel \mathbf{x} - P_{C_i}(\mathbf{x}_m)\parallel^2 - \parallel \mathbf{x}_m - P_{C_i}(\mathbf{x}_m)\parallel^2}{2\parallel \mathbf{x}_m - P_{C_i}(\mathbf{x}_m)\parallel}. \quad (8)$$

Summing these majorizations over *i* leads to quadratic majorization of *D*(**x**) and ultimately to the MM algorithm map

$$\psi(\mathbf{x}) = \arg\min_{\mathbf{z} \in S} \left\{ \frac{1}{2}\sum_{i=1}^{k} w_i \parallel \mathbf{z} - P_{C_i}(\mathbf{x})\parallel^2 \right\}$$

with weights $w_i = \gamma_i \|\mathbf{x} - P_{C_i}(\mathbf{x})\|^{-1}$. When the C_i are singletons and $S = \mathbb{R}^d$, the map $\psi(\mathbf{x})$ implements Weiszfeld's algorithm for solving the Fermat-Weber problem [28, 29].

The quadratic majorization of *D*(**x**) just derived can be rewritten as

$$g(\mathbf{x}|\mathbf{x}_m) = \frac{1}{2}\left(\sum_{i=1}^{k} w_i\right) \parallel \mathbf{x} - \sum_i \alpha_i P_{C_i}(\mathbf{x}_m)\parallel^2 + c,$$

where

$$\alpha_i = \frac{w_i}{\sum_{i=1}^{k} w_i},$$

depend on weights $w_i = \gamma_i \|\mathbf{x}_m - P_{C_i}(\mathbf{x}_m)\|^{-1}$ and *c* is a constant that does not depend on **x**. Thus, the MM update boils down to projection onto *S* of a convex combination of the projections of the previous iterate onto the sets C_i; in symbols we have

$$\mathbf{x}_{m+1} = P_S\left[\sum_i \alpha_i P_{C_i}(\mathbf{x}_m)\right]. \quad (9)$$

The majorization (8) involves dividing by 0 when $\mathbf{x}_m$ belongs to C_i. This singularity also bedevils Weiszfeld's algorithm. Fortunately, perturbation of the objective function salvages the situation. One simply replaces the function *D*(**x**) by the related function

$$D_\varepsilon(\mathbf{x}) = \sum_{j=1}^{k} \gamma_j \sqrt{d(\mathbf{x}, C_j)^2 + \varepsilon}$$

for ε small and positive. Ben-Tal and Teboulle [2] cover further examples of this perturbation strategy. In any case, observe that the smooth function $f_\varepsilon(u) = \sqrt{u^2 + \varepsilon}$ has derivatives

$$f_\varepsilon'(u) = \frac{u}{\sqrt{u^2 + \varepsilon}}, \; f_\varepsilon''(u) = \frac{\varepsilon}{(u^2 + \varepsilon)^{3/2}}$$

and is therefore strictly increasing and strictly convex on the interval $[0, \infty)$. Hence, the function $D_\varepsilon(\mathbf{x})$ is also convex. Because $\sqrt{u^2 + \varepsilon} - \sqrt{\varepsilon}$ is a good approximation to $u \; 0$, the solutions of the two problems should be close. In fact, we will show later that the minimum point of $D_\varepsilon(\mathbf{x})$ tends to the minimum point of $D(\mathbf{x})$ as ε tends to 0. In the presence of multiple minima, this claim must be rephrased in terms of cluster points.

The majorization $d(\mathbf{x}, C_j) \quad \|\mathbf{x} - P_{C_j}(\mathbf{x}_m)\|$ around the current iterate $\mathbf{x}_m$ yields the majorization

$$\sqrt{d(\mathbf{x}, C_j)^2 + \varepsilon} \leq \sqrt{\| \mathbf{x} - P_{C_j}(\mathbf{x}_m) \|^2 + \varepsilon}.$$

Application of the majorization (7) implies the further majorization

$$D_\varepsilon(\mathbf{x}) \leq \frac{1}{2} \sum_{j=1}^{k} \gamma_j \frac{\| \mathbf{x} - P_{C_j}(\mathbf{x}_m) \|^2}{\sqrt{\| \mathbf{x}_m - P_{C_j}(\mathbf{x}_m) \|^2 + \varepsilon}} + c,$$

where c is an irrelevant constant. The corresponding MM update $\mathbf{x}_{m+1}$ is identical to the previous MM update (9) except for one difference. The weights w_i are now defined by the benign formula

$$w_i = \frac{\gamma_i}{\sqrt{\| \mathbf{x}_m - P_{C_i}(\mathbf{x}_m) \|^2 + \varepsilon}}$$

involving no singularity.

5 Examples

We now consider four examples illustrating the performance of the MM algorithm and framing our expectations for convergence. The subgradient algorithm [22] serves as a benchmark for comparison throughout. This algorithm relies on the updates

$$\mathbf{x}_{m+1} = P_S \left[\mathbf{x}_m - \eta_m \sum_{i=1}^{k} \gamma_i \mathbf{v}_{im} \right],$$

where

$$\mathbf{v}_{im} = \begin{cases} \dfrac{\mathbf{x}_m - P_{C_i}(\mathbf{x}_m)}{d(\mathbf{x}_m, C_i)} & \text{if } \mathbf{x}_m \notin C_i \\ 0 & \text{if } \mathbf{x}_m \in C_i, \end{cases}$$

and the nonnegative constants η_m satisfy $\sum_{m=1}^{\infty} \eta_m = \infty$ and $\sum_{m=1}^{\infty} \eta_m^2 < \infty$. The weights γ_i equal 1 in all examples except the last.

5.1 Five Cubes and a Ball in $\mathbf{R}^3$

Our first example is taken from the reference [22]. This three-dimensional example involves five cubes C_i with side lengths equal to 2 and centers (0, −4, 0), (−4, 2, −3), (−3, −4, 2), (−5, 4, 4), and (−1, 8, 1). The set S is a ball with center (0, 2, 0) and radius 1. Iteration commences at the point $\mathbf{x}_1 = (0, 2, 0) \in S$ and takes subgradient steps with $\eta_m = 1/m$. Table 1 shows the MM iterates with $\varepsilon = 0$. Convergence to machine precision occurs within 30 iterations. In contrast, Table 2 shows that parameter values (x_1, x_2, x_3) are still changing after 10^6 subgradient iterates. For brevity we omit a second example of four squares and a disk in $\mathbb{R}^2$ from the same source [22]. In this example the superiority of the MM algorithm over the subgradient algorithm is equally evident.

5.2 The Closest Point to Three Disks in $\mathbf{R}^2$

This example from the reference [21] illustrates the advantage of minimizing a sequence of approximating functions $D_{\varepsilon_m}(\mathbf{x})$. The sets C_i are three unit balls in $\mathbb{R}^2$ centered at (0, 2), (2, 0), and (−2, 0). The set S equals $\mathbb{R}^2$. The minimum distance occurs at (0, 1) as can be easily verified by checking the optimality conditions spelled out in Proposition 4.3 in [21]. Figure 1 displays the iteration paths for 50 different starting values (dots) and their corresponding fixed point (the square). Along the mth leg of the path, we set ε_m to be max$\{10^{-m}, 10^{-16}\}$. The solution to the current problem is taken as the initial point for the next problem. All solution paths initially converge to a point just below (0,1) and then march collectively upwards to (0,1). The passage of the MM iterates through the unit balls is facilitated by our strategy of systematically reducing ε. Table 3 shows the subgradient and MM iterates starting from the point (5,7).

5.3 Three Collinear Disks in $\mathbf{R}^2$

Here we illustrate the behavior of the MM algorithm when there is more than one solution. Consider two unit balls in $\mathbb{R}^2$ centered at (2, 0), and (−2, 0), and take S to be the unit ball centered at the origin. There is a continuum of solutions extending along the line segment from (−1, 0) to (1, 0), as can be verified by the optimality conditions provided by Theorem 3.2 in [22]. Figure 2 shows the iteration paths for 100 different initial values (dots) and their corresponding fixed points (squares). In this example, we take $\varepsilon = 0$. Although the iterates

are not guaranteed to converge and may in principle cycle among multiple cluster points, this behavior is not observed in practice. The iterates simply converge to different fixed points depending on where they start. Table 4 compares the iterations for the subgradient method and the MM algorithm starting from the point (1.5,0.25). The two algorithms converge to different solution points, but at drastically different rates.

5.4 Kuhn's Problem

Our last example was originally concocted by Kuhn [14] to illustrate how Weiszfeld's algorithm can stall when its iterates enter one of the sets C_i. Although this event rarely occurs in practice, characterizing the initial conditions under which it happens has been a subject of intense scrutiny [5, 6, 7, 8, 15]. The occasional failure of Weiszfeld's algorithm prompted Vardi and Zhang [27] to redesign it. Their version preserves the descent property but differs substantially from ours. In any event the example shown in Figure 3 involves two points with weights γ_i proportional to 5 placed at (59,0) and (20,0) and two more points with weights proportional to 13 placed at (−20, 48) and (−20, −48). The optimal point is the origin. Starting at (44,0), Weiszfeld's algorithm stalls at (20,0) after one iteration. Our MM iterates (dots) with ε decreasing from 0.1 to 0, in contrast, move across (20.0) and correctly converge to (0,0) to within machine precision in 99 steps. Table 5 compares the progress achieved by the MM and subgradient methods. Note that when ε is 0.1, the MM algorithm overshoots the true answer and then comes back to (0, 0) after setting ε to be 0. The subgradient algorithm makes solid progress early but subsequently slows down on this almost smooth problem.

6 Convergence Theory

Before embarking on a proof of convergence, it is prudent to discuss whether a minimum point exists and is unique. Recall that a continuous function attains its minimum on a compact set. Thus, problem (6) possesses a minimum whenever S is bounded. If S is unbounded, then one can substitute boundedness of one or more of the sets C_i. In this circumstance, $D(\mathbf{x})$ is coercive in the sense that $\lim_{\|\mathbf{x}\|\to\infty} D(\mathbf{x}) = \infty$. As pointed out in Proposition 3.1 of the reference [22], coerciveness is sufficient to guarantee existence. Because $D(\mathbf{x})$ $D_\varepsilon(\mathbf{x})$, the perturbed criterion $D_\varepsilon(\mathbf{x})$ is coercive whenever the original criterion $D(\mathbf{x})$ is coercive. Henceforth, we will assume that S or at least one of the C_i is bounded.

A strictly convex function possesses at most one minimum point on a convex set. The function $|x|$ shows that this sufficient condition for uniqueness is hardly necessary. In the Fermat-Weber problem, where the closed convex sets $C_i = \{\mathbf{x}_i\}$ are singletons, the function $D(\mathbf{x})$ is strictly convex if and only if the points $\mathbf{x}_i$ are non-collinear. To generalize this result, we require the sets C_i to be non-collinear. Geometrically this says that it is impossible to draw a straight line that passes through all of the C_i. Non-collinearity can only be achieved when $k > 2$ and $\cap_{i=1}^{k} C_i = \varnothing$. We also require the C_i to be strictly convex. A set C is said to be strictly convex if the interior of the line segment $[\mathbf{x}, \mathbf{y}]$ connecting two different points $\mathbf{x}$ and $\mathbf{y}$ of C lies in the interior of C. Put another way, the boundary of C can contain no line segments. A singleton or a closed ball is strictly convex, but a closed box is not.

Proposition 6.1. *If the closed convex sets $C_1, \ldots, C_k$ are strictly convex but not collinear, then $D(\mathbf{x})$ is strictly convex.*

Proof. Suppose the contrary is true, and choose $\mathbf{x}$ $\mathbf{y}$ and α strictly between 0 and 1 so that

$$D[\alpha\mathbf{x}+(1-\alpha)\mathbf{y}]=\alpha D(\mathbf{x})+(1-\alpha)D(\mathbf{y}). \quad (10)$$

Let L be the line $\{s\mathbf{x}+(1-s)\mathbf{y} : s \in \mathbb{R}\}$ passing through the points $\mathbf{x}$ and $\mathbf{y}$. Then there exists at least one C_j such that $L \cap C_j = \varnothing$. In particular, $\mathbf{x}$, $\mathbf{y}$, and $\alpha\mathbf{x} + (1-\alpha)\mathbf{y}$ all fall outside this C_j. Equality (10) implies that

$$\alpha \parallel \mathbf{x} - P_{C_j}(\mathbf{x}) \parallel$$
$$+(1-\alpha) \parallel \mathbf{y}$$
$$- P_{C_j}(\mathbf{y}) \parallel$$
$$= \parallel \alpha\mathbf{x}$$
$$+(1-\alpha)\mathbf{y}$$
$$- P_{C_j}[\alpha\mathbf{x}+(1-\alpha)\mathbf{y}] \parallel \leq \parallel \alpha\mathbf{x}$$
$$+(1-\alpha)\mathbf{y} - \alpha P_{C_j}(\mathbf{x})$$
$$-(1-\alpha)P_{C_j}(\mathbf{y}) \parallel \leq \alpha \parallel \mathbf{x}$$
$$- P_{C_j}(\mathbf{x}) \parallel$$
$$+(1-\alpha) \parallel \mathbf{y}$$
$$- P_{C_j}(\mathbf{y}) \parallel .$$

Since the projection of a point onto C_j is unique, these sandwich inequalities entail

$$P_{C_j}[\alpha\mathbf{x}+(1-\alpha)\mathbf{y}]=\alpha P_{C_j}(\mathbf{x})+(1-\alpha)P_{C_j}(\mathbf{y}).$$

If $P_{C_j}(\mathbf{x})$ $P_{C_j}(\mathbf{y})$, then the strict convexity of C_j implies the convex combination $\alpha P_{C_j}(\mathbf{x}) +(1-\alpha)P_{C_j}(\mathbf{y})$ is interior to C_j. Hence, this point cannot be the closest point to the external point $\alpha\mathbf{x} + (1-\alpha)\mathbf{y}$. Therefore, consider the possibility $P_{C_j}(\mathbf{x}) = P_{C_j}(\mathbf{y}) = \mathbf{z}$. Equality can occur in the inequality

$$\parallel \alpha\mathbf{x}+(1-\alpha)\mathbf{y} - \mathbf{z} \parallel \leq \alpha \parallel \mathbf{x} - \mathbf{z} \parallel +(1-\alpha) \parallel \mathbf{y} - \mathbf{z} \parallel$$

only when $\mathbf{x} - \mathbf{z} = t(\mathbf{y} - \mathbf{z})$ for some t 1. This relation shows that

$$\mathbf{z}=\frac{1}{1-t}\mathbf{x} - \frac{t}{1-t}\mathbf{y}$$

belongs to $L \cap C_j$, contradicting our hypothesis. Thus, $D(\mathbf{x})$ is strictly convex.

The next result shows that the function $D_\varepsilon(\mathbf{x})$ inherits strict convexity from $D(\mathbf{x})$. Therefore, when $D(\mathbf{x})$ is strictly convex, $D_\varepsilon(\mathbf{x})$ possesses a unique minimum point.

Proposition 6.2. *If $D(\mathbf{x})$ is strictly convex, then $D_\varepsilon(\mathbf{x})$ is also strictly convex.*

Proof. Fix arbitrary $\mathbf{x} \neq \mathbf{y}$ and α strictly between 0 and 1. The strict convexity of $D(\mathbf{x})$ implies that there is at least one j such that

$$d(\alpha\mathbf{x}+(1-\alpha)\mathbf{y}, C_j) < \alpha d(\mathbf{x}, C_j)+(1-\alpha)d(\mathbf{y}, C_j).$$

The strict inequality

$$\sqrt{d(\alpha\mathbf{x}+(1-\alpha)\mathbf{y}, C_j)^2+\varepsilon} \quad < \sqrt{[\alpha d(\mathbf{x}, C_j)+(1-\alpha)d(\mathbf{y}, C_j)]^2+\varepsilon},$$
$$\leq \alpha\sqrt{d(\mathbf{x}, C_j)^2+\varepsilon}+(1-\alpha)\sqrt{d(\mathbf{y}, C_j)^2+\varepsilon},$$

follows because the function $f_\varepsilon(u)=\sqrt{u^2+\varepsilon}$ is a strictly increasing and convex. Summing over j gives the desired result.

We now clarify the relationship between the minima of the $D_\varepsilon(\mathbf{x})$ and $D(\mathbf{x})$ functions.

Proposition 6.3. *For a sequence of constants ε_m tending to 0, let $\mathbf{y}_m$ be a corresponding sequence minimizing $D_{\varepsilon_m}(\mathbf{x})$. If $\mathbf{y}$ is the unique minimum point of $D(\mathbf{x})$, then $\mathbf{y}_m$ tends to $\mathbf{y}$. If $D(\mathbf{x})$ has multiple minima, then every cluster point of the sequence $\mathbf{y}_m$ minimizes $D(\mathbf{x})$.*

Proof. To prove the assertion, consider the inequalities

$$D(\mathbf{y}_m) \leq D_{\varepsilon_m}(\mathbf{y}_m) \leq D_{\varepsilon_m}(\mathbf{x}) \leq D_1(\mathbf{x})$$

for any $\mathbf{x} \in S$ and $\varepsilon_m \leq 1$. Taking limits along the appropriate subsequences proves that the cluster points of the sequence $\mathbf{y}_m$ minimize $D(\mathbf{x})$. Convergence to a unique minimum point $\mathbf{y}$ occurs provided the sequence $\mathbf{y}_m$ is bounded. If S is bounded, then $\mathbf{y}_m$ is bounded by definition. On the other hand, if any C_j is bounded, then $D(\mathbf{x})$ is coercive, and the inequality $D(\mathbf{y}_m) \leq D_1(\mathbf{x})$ forces $\mathbf{y}_m$ to be bounded.

The convergence theory of MM algorithms hinges on the properties of the algorithm map $\psi(\mathbf{x}) \equiv \arg\min_\mathbf{y} g(\mathbf{y} \mid \mathbf{x})$. For easy reference, we state a simple version of Meyer's monotone convergence theorem [19] instrumental in proving convergence in our setting.

Proposition 6.4. *Let $f(\mathbf{x})$ be a continuous function on a domain S and $\psi(\mathbf{x})$ be a continuous algorithm map from S into S satisfying $f(\psi(\mathbf{x})) < f(\mathbf{x})$ for all $\mathbf{x} \in S$ with $\psi(\mathbf{x}) \neq \mathbf{x}$. Suppose for some initial point $\mathbf{x}_0$ that the set $\mathscr{L}_f(\mathbf{x}_0) \equiv \{\mathbf{x} \in S : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ is compact. Then (a) all cluster points are fixed points of $\psi(\mathbf{x})$, and (b) $\lim_{m\to\infty}\|\mathbf{x}_{m+1} - \mathbf{x}_m\| = 0$.*

Note that Proposition 6.4 also ensures the existence of at least one cluster point for the sequence of iterates $\mathbf{x}_{m+1} = \psi(\mathbf{x}_m)$. Additionally, the convergence of the MM iterates (9) to a stationary point of $f(\mathbf{x})$ follows immediately provided the fixed points of $\psi(\mathbf{x})$ are stationary points of $f(\mathbf{x})$ and $\psi(\mathbf{x})$ possesses only finitely many fixed points.

Let us verify the conditions of Proposition 6.4 for minimizing $D_\varepsilon(\mathbf{x})$. The function $D_\varepsilon(\mathbf{x})$ is continuous on its domain S, and the set $\mathscr{L}_{D_\varepsilon}(\mathbf{x}_0)$ is compact for any initial point $\mathbf{x}_0$ since either S is compact or $D_\varepsilon(\mathbf{x})$ is coercive. The continuity of the algorithm map follows immediately from the continuity of the projection mapping. Finally, we need to prove that $D_\varepsilon(\psi(\mathbf{x})) < D_\varepsilon(\mathbf{x})$ whenever $\mathbf{x} \neq \psi(\mathbf{x})$. First observe that $\psi(\mathbf{x}) = \mathbf{x}$ if and only if the MM surrogate function satisfies $g_\varepsilon(\mathbf{x} \mid \mathbf{x}) = \min_\mathbf{y} g_\varepsilon(\mathbf{y} \mid \mathbf{x})$. Since $g_\varepsilon(\mathbf{y} \mid \mathbf{x})$ has a unique minimizer, we have the strict inequality $g_\varepsilon(\psi(\mathbf{x}) \mid \mathbf{x}) < g_\varepsilon(\mathbf{x} \mid \mathbf{x})$ whenever $\mathbf{x}$ is not a fixed point of ψ. This forces a decrease in the objective function $D_\varepsilon(\mathbf{x})$ and makes the MM algorithm strictly monotone outside the set of stationary points.

We now argue that the fixed points of the algorithm map $\psi(\mathbf{x})$ are stationary points of $D_\varepsilon(\mathbf{x})$. We will show, in fact, that the two sets of points coincide. To accomplish this, we need to determine the gradients of $D_\varepsilon(\mathbf{x})$ and $g_\varepsilon(\mathbf{x} \mid \mathbf{y})$. Recall that $f_\varepsilon(u)$ is strictly increasing and strictly convex. As a consequence the functions $f_\varepsilon(\|\mathbf{x}\|)$ and $f_\varepsilon[d(\mathbf{x}, C_j)]$ are convex. Even more remarkable is the fact that both functions are continuously differentiable. When $\mathbf{x} \neq \mathbf{0}$, the function $\|\mathbf{x}\|$ is differentiable. Likewise, when $\mathbf{x} \notin C_j$, the function $d(\mathbf{x}, C_j)$ is differentiable. Therefore, the chain rule implies

$$\nabla f_\varepsilon(\|\mathbf{x}\|) = \frac{\|\mathbf{x}\|}{\sqrt{\|\mathbf{x}\|^2 + \varepsilon}} \frac{\mathbf{x}}{\|\mathbf{x}\|} = \frac{\mathbf{x}}{\sqrt{\|\mathbf{x}\|^2 + \varepsilon}} \quad (11)$$

$$\nabla f_\varepsilon[d(\mathbf{x}, C_j)] = \frac{d(\mathbf{x}, C_j)}{\sqrt{d(\mathbf{x}, C_j)^2 + \varepsilon}} \frac{\mathbf{x} - P_{C_j}(\mathbf{x})}{d(\mathbf{x}, C_j)} = \frac{\mathbf{x} - P_{C_j}(\mathbf{x})}{\sqrt{d(\mathbf{x}, C_j)^2 + \varepsilon}}, \quad (12)$$

respectively.

By continuity one expects the gradients to be defined for $\mathbf{x} = \mathbf{0}$ and $\mathbf{x} \in C_j$ by the corresponding limit of $\mathbf{0}$. In the former case the expansion

$$\sqrt{\|\mathbf{x}\|^2 + \varepsilon} - \sqrt{\varepsilon} = \sqrt{\varepsilon}\sqrt{1 + \frac{\|\mathbf{x}\|^2}{\varepsilon}} - \sqrt{\varepsilon} = \frac{1}{2}\frac{\|\mathbf{x}\|^2}{\sqrt{\varepsilon}} + \sqrt{\varepsilon}\, o\left(\frac{\|\mathbf{x}\|^2}{\varepsilon}\right)$$

shows that $\nabla f_\varepsilon(\|\mathbf{0}\|) = \mathbf{0}$. In the latter case the expansion

$$\sqrt{d(\mathbf{y}, C_j)^2 + \varepsilon} - \sqrt{\varepsilon} = \frac{1}{2}\frac{d(\mathbf{y}, C_j)^2}{\sqrt{\varepsilon}} + \sqrt{\varepsilon}\, o\left[\frac{d(\mathbf{y}, C_j)^2}{\varepsilon}\right]$$

and the bound $d(\mathbf{y}, C_j) = |d(\mathbf{y}, C_j) - d(\mathbf{x}, C_j)| \leq \|\mathbf{y} - \mathbf{x}\|$ for $\mathbf{x} \in C_j$ likewise show that $\nabla f_\varepsilon[d(\mathbf{x}, C_j)] = \mathbf{0}$. Consequently, equations (11) and (12) hold for all $\mathbf{x} \in \mathbb{R}^d$. It follows that both $D_\varepsilon(\mathbf{x})$ and $g_\varepsilon(\mathbf{x} \mid \mathbf{y})$ are differentiable on $\mathbb{R}^d$, with gradients

$$\nabla D_\varepsilon(\mathbf{x}) = \sum_{j=1}^{k} \gamma_j \frac{\mathbf{x} - P_{C_j}(\mathbf{x})}{\sqrt{d(\mathbf{x}, C_j)^2 + \varepsilon}},$$

and

$$\nabla g_\varepsilon(\mathbf{x}|\mathbf{y}) = \sum_{j=1}^{k} \gamma_j \frac{\mathbf{x} - P_{C_j}(\mathbf{y})}{\sqrt{d(\mathbf{y}, C_j)^2 + \varepsilon}}, \quad (13)$$

respectively. Note that $\mathbf{y} \in S$ minimizes $D_\varepsilon(\mathbf{x})$ over S if and only if

$$\sum_{j=1}^{k} \gamma_j \frac{\langle \mathbf{y} - P_{C_j}(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle}{\sqrt{d(\mathbf{y}, C_j)^2 + \varepsilon}} \geq 0,$$

for all $\mathbf{x} \in S$. This inequality, however, is equivalent to the inequality $\langle \nabla g_\varepsilon(\mathbf{y}\,|\,\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$ 0, for all $\mathbf{x} \in S$, which in turn holds if and only if $\mathbf{y}$ is a fixed point of $\psi(\mathbf{x})$. If $D(\mathbf{x})$ is strictly convex, then $D_\varepsilon(\mathbf{x})$ has a unique minimum point, and $\psi(\mathbf{x})$ has exactly one fixed point.

Thus, Proposition 6.3 and Proposition 6.4 together tell us that $\mathbf{y}$ is a solution to (6) if there is a sequence of ε_m tending to zero and a sequence of points $\mathbf{y}_m$ tending to $\mathbf{y}$ that satisfy

$$\left\langle -\sum_{j=1}^{k} \gamma_j \frac{\mathbf{y}_m - P_{C_j}(\mathbf{y}_m)}{\sqrt{d(\mathbf{y}_m, C_j)^2 + \varepsilon_m}}, \mathbf{x} - \mathbf{y}_m \right\rangle \leq 0, \quad (14)$$

for all $\mathbf{x} \in S$. The above sufficient condition becomes necessary as well if $D(\mathbf{x})$ is strictly convex. As a sanity check, when the sets $S \cap C_j$ are all empty and the weights γ_j are identical, we recover the characterization of the optimal points given in Theorem 3.2 of reference [22], albeit under the more restrictive assumption of strict convexity.

7 Conclusion

There is admittedly an art to applying the MM principle. The majorization presented here is specific to Euclidean distances, and changing the underlying norm would require radical revision. Nonetheless, when the MM principle applies, the corresponding MM algorithm can be effective, simple to code, and intuitively appealing. Here the principle lit the way to an efficient numerical algorithm for solving the Euclidean version of the generalized Heron problem using only elementary principles of smooth convex analysis. We also suggested a simple yet accurate approximation of the problem that removes the singularities of the MM algorithm and Weiszfeld's earlier algorithm. Similar advantages accrue across a broad spectrum of optimization problems. The ability of MM algorithms to handle high-dimensional problems in imaging, genomics, statistics, and a host of other fields testifies to the potency of a simple idea consistently invoked. Mathematical scientists are well advised to be on the lookout for new applications.

Acknowledgments

References

1. Becker MP, Yang I, Lange K. EM algorithms without missing data. Stat. Methods Med. Res. 1997; 6:38–54. [PubMed: 9185289]

2. Ben-Tal, A.; Teboulle, M. Optimization, Edited by S. Dolecki, Lecture Notes in Mathematics. Vol. Vol. 1405. Berlin / Heidelberg: Springer; 1989. A smoothing technique for nondifferentiable optimization problems; p. 1-11.

3. Bertsekas, DP. Convex Optimization Theory. Belmont, MA: Athena Scientific; 2009.

4. Borwein, JM.; Lewis, AS. Convex Analysis and Nonlinear Optimization: Theory and Examples. New York: Springer; 2000.

5. Brimberg J. The Fermat-Weber location problem revisited. Math. Program. 1995; 71:71–76.

6. Brimberg J. Further notes on convergence of the Weiszfeld algorithm. Yugosl. J. Oper. Res. 2003; 13:199–206.

7. Cánovas L, Marín A, Cañavate R. On the convergence of the Weiszfeld algorithm, Math. Program. 2002; 93:327–330.

8. Chandrasekaran R, Tamir A. Open questions concerning Weiszfeld's algorithm for the Fermat-Weber location problem. Math. Program. 1989; 44:293–295.

9. Courant, R.; Robbins, H. What is Mathematics?. New York: Oxford University Press; 1961.

10. Duchi, J.; Shalev-Shwartz, S.; Singer, Y.; Chandra, T. Efficient projections onto the ℓ_1-ball for learning in high dimensions. 25th International Conference on Machine Learning – Helsinki 2008; Proc. Int. Conf. Mach. Learn., The International Machine Learning Society; 2008. p. 272-279.

11. Gueron S, Tessler R. The Fermat-Steiner problem. Amer. Math. Monthly. 2002; 109:443–451.

12. Hiriart-Urruty, J-B.; Lemaréchal, C. Fundamentals of Convex Analysis. Berlin: Springer-Verlag; 2001.

13. Krarup J, Vajda S. On Torricelli's geometrical solution to a problem of Fermat. IMA J. Math. Appl. Bus. Indust. 1997; 8:215–224.

14. Kuhn, HW. On a pair of dual nonlinear programs, in Nonlinear Programming. Abadie, J., editor. Amsterdam: North-Holland Publishing; 1967. p. 37-54.

15. Kuhn HW. A note on Fermat's problem. Math. Program. 1973; 4:98–107.

16. Lange, K. Numerical Analysis for Statisticians. second edition. New York: Springer; 2010.

17. Lange K, Hunter DR, Yang I. Optimization transfer using surrogate objective functions (with discussion. J. Comput. Graph. Statist. 2000; 9:1–20.

18. Love, RF.; Morris, JG.; Wesolowsky, GO. Facilities Location: Models and Methods, Publications in Operations Research Series. Vol. Vol. 7. Amsterdam: North-Holland Publishing; 1988.

19. Meyer R. Sufficient conditions for the convergence of monotonic mathematical programming algorithms. J. Comput. System Sci. 1976; 12:108–121.

20. Michelot C. A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$. J. Optim. Theory Appl. 1986; 50:195–200.

21. Mordukhovich B, Nam NM. Applications of variational analysis to a generalized Fermat-Torricelli problem. J. Optim. Theory Appl. 2011; 148:431–454.

22. Mordukhovich B, Nam NM, Salinas J. Solving a generalized Heron problem by means of convex analysis. Amer. Math. Monthly. 2012; 119:87–99.

23. Mordukhovich BS, Nam NM, Salinas J. Applications of variational analysis to a generalized Heron problems. Appl. Anal. 2011:1–28.

24. Ortega, JM.; Rheinboldt, WC. Iterative Solutions of Nonlinear Equations in Several Variables. New York: Academic; 1970.

25. Rockafellar, RT. Convex Analysis. Princeton, NJ: Princeton University Press; 1996.

26. Ruszczy ski, A. Nonlinear Optimization. Princeton, NJ: Princeton University Press; 2006.

27. Vardi Y, Zhang C-H. A modified Weiszfeld algorithm for the Fermat-Weber location problem. Math. Program. 2001; 90:559–566.

28. Weiszfeld E. Sur le point pour lequel la somme des distances de *n* points donnés est minimum. Tôhoku Mathematics Journal. 1937; 43:355–386.

29. Weiszfeld E, Plastria F. On the point for which the sum of the distances to *n* given points is minimum. Ann. Oper. Res. 2009; 67:7–41.

30. Wesolowsky GO. The Weber problem: Its history and perspectives. Location Science . 1993; 1:5–23.

Biographies

Eric C. Chi received his Ph.D. from Rice University in 2011 under the direction of David W. Scott. He is currently a postdoctoral fellow at the University of California, Los Angeles.

Kenneth Lange is the Rosenfeld Professor of Computational Genetics at UCLA. He is also Chair of the Department of Human Genetics and Professor of Biomathematics and Statistics. He has held permanent or visiting appointments at the University of New Hampshire, University of Michigan, University of Helsinki, M.I.T, Harvard, and Stanford. He has authored four advanced textbooks and published over 225 scientific papers.
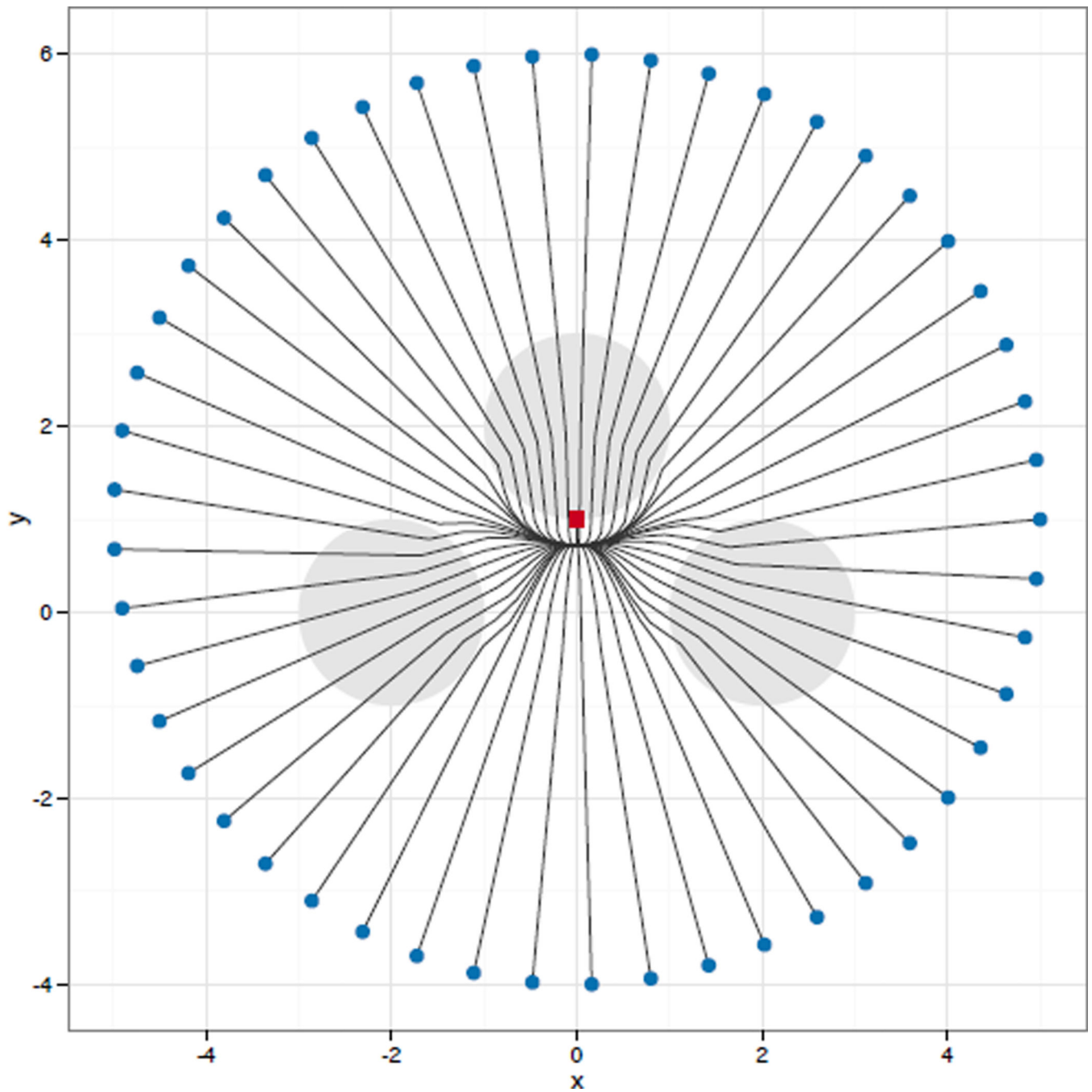
Figure 1.
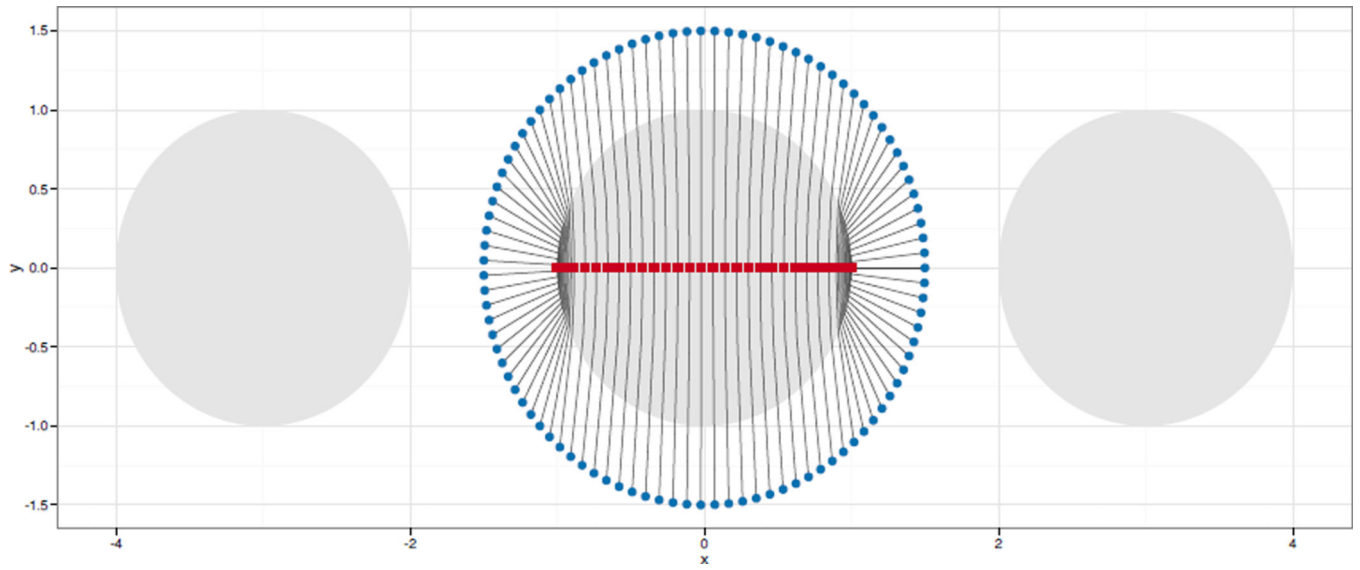Finding the closest point to three disks in $\mathbb{R}^2$.

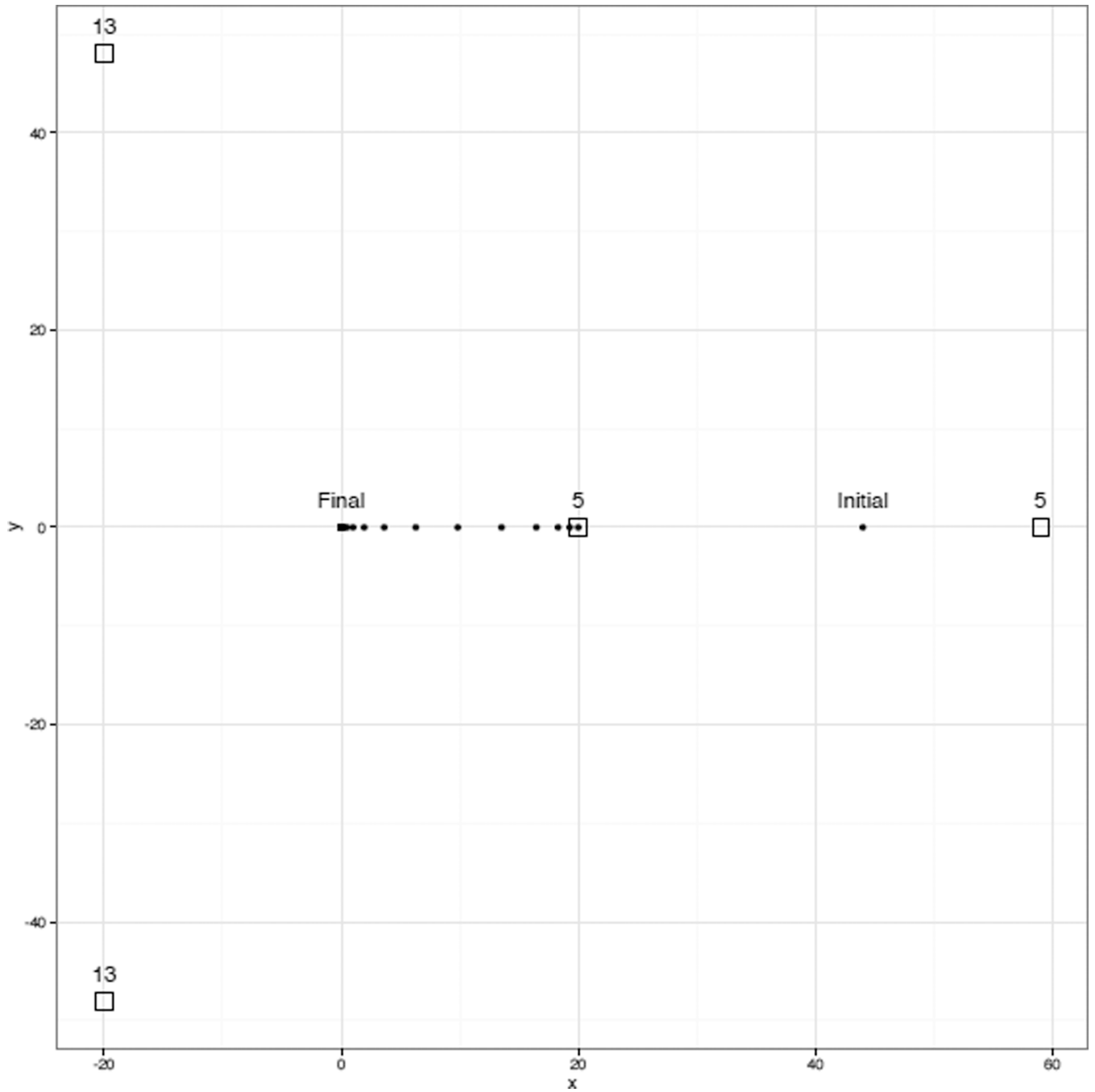**Figure 2.**
An example with a continuum of solutions.

Figure 3.
A problem where Weiszfeld's algorithm fails to converge.

Table 1

Cubes and ball example in $\mathbb{R}^3$: the MM Algorithm.

Iteration	x_1	x_2	x_3
1	0.00000000000000	2.00000000000000	0.00000000000000
2	−0.93546738305698	1.66164748416805	0.10207032020482
3	−0.92881282698649	1.63915389878166	0.08424264751830
4	−0.92645373003448	1.63220797263449	0.08007815377225
5	−0.92567602259658	1.63004821970935	0.07911751670489
6	−0.92542515217106	1.62937435413374	0.07889815178685
7	−0.92534495711879	1.62916364685109	0.07884864943702
8	−0.92531944712805	1.62909766226627	0.07883765997470
9	−0.92531135783449	1.62907697582185	0.07883527888603
10	−0.92530879826106	1.62907048520349	0.07883478238381
20	−0.92530761702316	1.62906751412014	0.07883466748783
30	−0.92530761701184	1.62906751409212	0.07883466748878
50	−0.92530761701184	1.62906751409212	0.07883466748878

Table 2

Cubes and ball example in $\mathbb{R}^3$: the Subgradient Algorithm.

Iteration	x_1	x_2	x_3
1	0.00000000000000	2.00000000000000	0.00000000000000
10	−0.92583298353433	1.63051788239768	0.07947484741743
100	−0.92531325048300	1.62908232435160	0.07883822912883
1000	−0.92530767419684	1.62906766065418	0.07883468589312
10000	−0.92530761758555	1.62906751554109	0.07883466757273
100000	−0.92530761701755	1.62906751410641	0.07883466748904
1000000	−0.92530761701233	1.62906751409334	0.07883466748881
1500000	−0.92530761701231	1.62906751409328	0.07883466748881
2000000	−0.92530761701229	1.62906751409324	0.07883466748881

Table 3

Three disks example in $\mathbb{R}^2$ starting from (5,7).

| | Subgradient Algorithm | | | MM Algorithm | |
Iteration	x_1	x_2	Iteration	x_1	x_2
10	0.7092649	1.2369866	10	0.2674080	0.7570688
100	0.0558764	0.9973310	100	0.0000000	0.7249706
1,000	0.0046862	0.9993844	1,000	0.0000000	0.9998002
10,000	0.0003955	0.9999274	1,800	0.0000000	0.9999999
100,000	0.0000334	0.9999957	1,850	0.0000000	1.0000000
1,000,000	0.0000028	0.9999998	1,900	0.0000000	1.0000000

Table 4

Three collinear disks example in $\mathbb{R}^2$ starting from (1.5, 0.25)

| | Subgradient Algorithm | | | MM Algorithm | |
Iteration	x_1	x_2	Iteration	x_1	x_2
10,000	0.9997648	0.0000223	10	0.9941149	0.0001308
100,000	0.9997648	0.0000040	20	0.9941149	0.0000000
1,000,000	0.9997648	0.0000007	30	0.9941149	0.0000000

Table 5

Kuhn's problem

	Subgradient Algorithm			MM Algorithm		
Iteration	x_1	x_2	**Iteration**	x_1	x_2
10	8.6984831	0.0000000	10	1.9448925	0.0000000
1,000	1.2966354	0.0000000	30	−0.0011998	0.0000000
100,000	0.1845171	0.0000000	60	−0.0012011	0.0000000
10,000,000	0.0259854	0.0000000	90	0.0000000	0.0000000