Clustering and the perturbed spatial median

Tuomo Valkonen **Tommi Kärkkäinen**

December 27, 2014

Abstract
The diff-convex (DC) problem of perturbed spatial median and the Weiszfeld algorithm in a framework for incomplete data is studied, and some level set theorems for general DC problems are provided. These results are then applied to study certain multiobjective formulations of clustering problems, and to yield a new algorithm for solving the multisource Weber problem.

Keywords: Weiszfeld algorithm, diff-convex, multisource Weber problem, clustering, multiobjective

Mathematics subject classification: 90C26, 90B85

1 Introduction

In this paper, we are interested in minimisation problems, where the objective function can be modelled as a perturbed version of the spatial median objective $\sum_{k=1}^{n} d(a_k, \cdot)$. More specifically, what concerns us are problems of the form

$$\min_p \left(\sum_{k=1}^{n} d(a_k, p) - \nu(p) \right) \tag{1}$$

for some fixed points $a_1, \ldots, a_n$, a convex function ν, and a distance d generalising the Euclidean metric.

On the application side, we are concerned with the general class of problems of locating one or more points $p_1, \ldots, p_s$ according to some optimality criterion involving another set of n fixed points and combinations of distances between all these points. In the single-prototype case ($s = 1$), popular objectives are the data means and the above-mentioned spatial median. In the latter case, the problem itself is then also known as the (Fermat-)Weber problem, and the Weiszfeld algorithm may be used to look for a solution [31, 16]. Multi-prototype ($s > 1$) variants of the location problem often somehow involve the single-facility case. In particular, in case of criteria of the K-means type [10, 25], the goal is to assign each vertex to the closest prototype p_j, with the prototypes being the data means, spatial medians, or other points somehow descriptive of the centres of the corresponding clusters. For an overview of work on this and other clustering problems, as well as a unifying framework for smoothed and approximating problems, we point the reader to [26].

Problem (1) seems at a first glance, to only involve a single prototype. However, we will work with a model for incomplete data that will let us model multi-prototype problems as single-prototype ones. It then turns out that the classic multisource Weber problem – the problem of finding the *K-spatial-medians* – otherwise also known as the location-allocation problem [9], fits this model, as well as do some of our own multi-criteria formulations of the clustering problem, among others.

Indeed, the algorithm we develop for (1), will be a further extension of the above-mentioned generalisation to incomplete data sets of the Weiszfeld algorithm in [17, 18, 28, 29]. This algorithm in its basic form [31, 16] seeks a minimiser to $\sum_{i=1}^{n} w_i d(a_i, \cdot)$ for the Euclidean distance in $\mathbb{R}^m$ by iterating

$$T : p \mapsto \frac{\sum_{i=1}^{n} s_i a_i}{\sum_{i=1}^{n} s_i} \quad \text{with } s_i = w_i / d(a_i, p). \tag{2}$$

Since the objective in (1) is a difference of convex functions (DC function; see [27, 15]), it is generally not convex. Therefore, being a local algorithm, our convergence results are weaker than in the above conventional case. The incomplete data sets also bring their own considerable problems, even under rather strict assumptions. In practise the results seem promising, however.

This paper is organised as follows. First, in Section 2 we elaborate the generalised Weiszfeld algorithm for problems of the form (1), and consider its convergence to what we will call *semi-critical* points, extending the results in [28, 29]. In Sections 3 and 4 we then study the application of our algorithm to solving some multi-objective formulations of the clustering problems and the K-spatial-medians, respectively. Finally, some experiments with these problems are presented in Section 5, and the paper is concluded in Section 6. Furthermore, Appendix A is devoted to proving some general level set theorems for DC functions that are useful for verifying the applicability of the algorithm. Moreover, some results relevant to the implementation of the described algorithm for more complex data sets, are presented in Appendix B.

2 The perturbed spatial median

Throughout most of this paper, we work with $n \geq 1$ vertices $a_1, \dots, a_n \in \mathbb{R}^m$ and diagonal positive-semidefinite matrices $W_1, \dots, W_n \in \mathbb{R}^{m \times m}$. The matrices W_i model the importance and incompleteness of the data, and typically have the form $W_i = w_i \rho_i$, for a weight $w_i > 0$ and a zero-one diagonal matrix ρ_i. A zero diagonal element of ρ_i indicates that the corresponding field of a_i is "missing", and an element with value one indicates that it is present. We assume (without loss of generality) that the data covers the whole space, i.e. $\sum_{i=1}^{n} \mathcal{R}(W_i) = \mathbb{R}^m$, with $\mathcal{R}$ denoting the range. The identity matrix is denoted by I.

With $\|\cdot\|$ denoting the Euclidean norm in $\mathbb{R}^m$, we now define the semi-norms and distance functions $d_i(p) \triangleq \|a_i - p\|_i \triangleq \|W_i(a_i - p)\|$, as well and the sum of distances function $f : \mathbb{R}^m \to \mathbb{R}$ as

$$f(p) \triangleq \sum_{i=1}^{n} d_i(p) = \sum_{i=1}^{n} \|W_i(a_i - p)\| . \tag{3}$$

A minimiser of f is called a *spatial median* of the points $\{a_i\}$. Existence and uniqueness in case of non-collinear data covering the whole space, follows as in [28, Theorem

3.1], where the problem (3) was studied under a more elaborate model for missing data.

Now, consider the problem of finding the minimum of (3) perturbed with the negation of a finite-valued convex function ν. That is, calling the objective function $f_\nu \triangleq f - \nu$, we consider the problem

$$\min_{p\in\mathbb{R}^m} f_\nu(p) = \min_{p\in\mathbb{R}^m} \sum_{i=1}^n d_i(p) - \nu(p). \tag{4}$$

Any solution of problem (4) will be called a *perturbed spatial median*. It turns out that a slightly modified Weiszfeld algorithm is still applicable for finding what we will call *semi-* and more generally *$\mathcal{D}$-critical points*, on the assumption that the subdifferentials of ν are in some sense properly contained in the range of the subdifferentials of $\sum_{i=1}^n d_i$ or if we can otherwise guarantee some boundedness properties.

For now we will, however, only require that ν is finite-valued. Then it will also have non-empty locally uniformly bounded subdifferentials, by e.g. [23, Corollary 24.5.1]. Recall that a set-valued mapping $F : X \rightrightarrows Y$ between metric spaces X and Y is *locally uniformly bounded at* $x \in X$ if there exists a neighbourhood U of x such that $\bigcup_{x'\in U} F(x')$ is bounded in Y.

2.1 Directions of descent

Notice that, since ν is convex, if we replace it with a linearisation $\tilde{\nu}_p^v(p') \triangleq \nu(p) + v^T(p'-p)$ for $v \in \partial\nu(p)$, then $-\nu \le -\tilde{\nu}_p^v$ and, furthermore, f_ν is dominated by the *upper convexification* $f_{\nu_p^v}$. Therefore, for any $p' \in \mathbb{R}^m$, for which $f(p') - \tilde{\nu}_p^v(p') < f(p) - \tilde{\nu}_p^v(p) = f(p) - \nu(p)$ it follows that $f(p') - \nu(p') < f(p) - \nu(p)$. This means that if some upper convexification at p is descending to some direction, so is f_ν itself.

The next theorem provides a sufficient condition for search direction and step length for the minimisation of f_ν. To state it, we need to introduce some notation. We write $\pi(p) \triangleq \{i \mid W_i(a_i - p) = 0\}$. The gradient of the differentiable components of f at p is then given by

$$g_\pi(p) \triangleq \sum_{i\notin\pi(p)} W_i^2 \frac{p-a_i}{\|p-a_i\|_i} = \sum_{i\notin\pi(p)} S_i(p)(p-a_i),$$

for $S_i(p) \triangleq W_i^2/d_i(p)$. We also define $S_\pi(p) \triangleq \sum_{i\notin\pi(p)} S_i(p)$, and the pseudoinverse of the (diagonal positive-semidefinite) matrix $S_\pi(p)$ as $S_\pi^\dagger(p)$. The orthogonal projection matrix into $\sum_{k\in\pi(p)} \mathcal{R}(W_k)$ is denoted $\rho_{\pi(p)}$, and the projection into the orthogonal complement as $\bar{\rho}_{\pi(p)}$. Let us also abbreviate $g_\pi^v(p) \triangleq g_\pi(p) - v$, and define

$$h(z, v; p) \triangleq g_\pi^v(p)^T z + \sum_{k\in\pi(p)} \|z\|_k \,.$$

Theorem 1. *Suppose $\nu(p) = v^T p$ for some $v \in \mathbb{R}^m$, and let $z \in \mathbb{R}^m$. Then $f_\nu(p + \omega z) < f_\nu(p)$, if $\omega \in (0, \Omega)$ with $\Omega \triangleq \Omega(p, v, z)$ defined as the supremum of ω' satisfying*

$$\omega'(z^T S_\pi(p) z) < -2h(z, v; p). \tag{5}$$

Additionally, there exists $z \neq 0$ with $\Omega(p, v, z) > 0$ if and only if there exists a direction of descent of f_ν at p.

Proof. We will write $p' \triangleq p + \omega z$, $g_\pi \triangleq g_\pi(p)$ and $\pi \triangleq \pi(p)$ to make the equations more legible.

Write

$$f(p) = \sum_{i \notin \pi} \frac{d_i(p)^2}{d_i(p)} \quad \text{and} \quad f(p') = \sum_{i \notin \pi} \frac{d_i(p')d_i(p)}{d_i(p)} + \sum_{k \in \pi} d_k(p').$$

As $d_i(p')d_i(p) - d_i(p)^2 = \frac{1}{2}\big(d_i(p')^2 - d_i(p)^2 - (d_i(p) - d_i(p'))^2\big)$, we have that

$$2(f_\nu(p') - f_\nu(p)) = \sum_{i \notin \pi} \frac{d_i(p')^2}{d_i(p)} - \sum_{i \notin \pi} \frac{d_i(p)^2}{d_i(p)} + \sum_{k \in \pi} 2d_k(p') - 2v^T(p' - p) - C,$$

where $C \triangleq \sum_{i \notin \pi}(d_i(p) - d_i(p'))^2/d_i(p)$ is non-negative. Using $p' = p + \omega z$ gives

$$d_i(p')^2 = \|(p - a_i) + \omega z\|_i^2 = d_i(p)^2 + 2\omega z^T W_i^2(p - a_i) + \omega^2 z^T W_i^2 z.$$

Thus, because $d_k(p') = d_k(p+\omega z) = \omega \|z\|_k$ for $k \in \pi$, we have that $f_\nu(p') - f_\nu(p) < 0$ holds if

$$2\omega z^T \left(\sum_{i \notin \pi} W_i^2 \frac{p - a_i}{d_i(p)} \right) + \omega^2 \sum_{i \notin \pi} \frac{z^T W_i^2 z}{d_i(p)} + 2\omega \sum_{k \in \pi} \|z\|_k - 2\omega v^T z < 0,$$

or, more compactly put,

$$\omega(z^T S_\pi z) < 2(-(g_\pi - v)^T z - \sum_{k \in \pi} \|z\|_k),$$

which gives the condition (5).

The second claim follows since, in fact, $h(z, v; p)$ is the directional derivative $f'_\nu(p; z)$. □

The next result provides further detail on calculating a step z. To specify it, we use the notation

$$Z(p) \triangleq \{z \in \mathbb{R}^m \mid \|z\| = 1, \bar{\rho}_{\pi(p)} z = 0\} \cup \{0\}$$

for the set of search directions in the subspace $\mathcal{R}(\rho_{\pi(p)}) = \sum_{k \in \pi(p)} \mathcal{R}(W_k)$ spanned by the non-differentiable components of f.

Lemma 1. *Let v and f_ν be as in Theorem 1. Let $\tilde{z} \in Z(p)$ be such that $h(\tilde{z}, v; p) < 0$ if such a choice exists. Otherwise choose $\tilde{z} = 0$. Suppose $\omega \in (0, 2)$ and that $p \in \mathbb{R}^m$ is not a minimiser of f_ν. Then*

$$z = z(p, v) \triangleq -\bar{\rho}_{\pi(p)} S_\pi^\dagger(p) g_\pi^v(p) + \alpha \tilde{z} \tag{6}$$

is a direction of descent for f_ν when $\alpha \in (0, \alpha_0)$, where $\alpha_0 \triangleq \alpha_0(\omega, \tilde{z}, v; p) > 0$ is the supremum of α for which z satisfies (5) at p for ω and v. Furthermore, $\alpha_0(2, \tilde{z}, v; p)$ gives for any $\omega \in (0, 2)$ a lower bound $\alpha_2(\tilde{z}, v; p) \in (0, \alpha_0]$ (strict if $-\bar{\rho}_{\pi(p)} g_\pi^v(p) \neq 0$), obtained as the supremum of α satisfying

$$\alpha \tilde{z}^T \rho_{\pi(p)} S_\pi(p) \tilde{z} \leq -h(\tilde{z}, v; p). \tag{7}$$

Proof. We will abbreviate $z \triangleq z(p,v)$, $g_\pi^v \triangleq g_\pi^v(p)$, $S_\pi \triangleq S_\pi(p)$, and $\pi = \pi(p)$ for legibility.

Inserting (6) into condition (5) of Theorem 1, we get

$$\omega((g_\pi^v)^T \bar\rho_\pi S_\pi^\dagger S_\pi S_\pi^\dagger \bar\rho_\pi g_\pi^v + \alpha^2 \tilde z^T \rho_\pi S_\pi \rho_\pi \tilde z - 2\alpha \tilde z^T \rho_\pi S_\pi \bar\rho_\pi S_\pi^\dagger g_\pi^v) < -2(-(g_\pi^v)^T \bar\rho_\pi S_\pi^\dagger g_\pi^v + \alpha (g_\pi^v)^T \rho_\pi \tilde z + \sum_{k\in\pi} \|\alpha \tilde z\|_k),$$

because $W_k z = \alpha W_k \tilde z$ for $k \in \pi$. As also $\rho_\pi S_\pi \bar\rho_\pi = 0$ and $S_\pi^\dagger S_\pi S_\pi^\dagger = S_\pi^\dagger$, this reduces to

$$\omega((g_\pi^v)^T \bar\rho_\pi S_\pi^\dagger g_\pi^v + \alpha^2 \tilde z^T \rho_\pi S_\pi \tilde z) - 2((g_\pi^v)^T \bar\rho_\pi S_\pi^\dagger g_\pi^v - \alpha h(\tilde z, v; p)) < 0, \tag{8}$$

where α has been taken outside norms because it is non-negative by assumption.

If $\tilde z = 0$, then α does not contribute to (8), so its choice is irrelevant and α_0 infinite. If, furthermore, $\bar\rho_\pi g_\pi^v = 0$, then $\min h(z, v; p) = 0$ over $\|z\| = 1$, and therefore by Theorem 1, p is a minimiser of f_ν. If, on the other hand, $\bar\rho_\pi g_\pi^v \neq 0$, then any $\omega < 2$ is valid.

If $\rho_\pi S_\pi \tilde z = 0$ but $\tilde z \neq 0$, then since $h(\tilde z, v; p) < 0$, we see that α can still be arbitrarily large, and any $\omega \in (0,2)$ is valid even for small α.

Suppose then that all the terms in (8) involving $\tilde z$ are non-zero. Whenever $0 < \omega < 2$, the inequality is either satisfied for $\alpha = 0$, or becomes an equality. Therefore, because the inequality is quadratic in α with the multiplier of the second-order term positive, and that of the first order term negative, there is for any $0 < \omega < 2$ an $\alpha_0(\omega, \tilde z, v; p) > 0$, such that $\alpha \in (0, \alpha_0(\omega, p, v))$ satisfies the inequality.

Setting $\omega = 2$ in (8), gives the condition for α_2. Furthermore, if α_2 satisfies (8) for $\omega = 2$, possibly non-strictly, it must continue to do so for $\omega < 2$, strictly if $(g_\pi^v)^T \bar\rho_\pi S_\pi^\dagger g_\pi^v \neq 0$ (which is equivalent to the condition in the statement). The lower bound on α_0 follows. □

Example 1.

i) When $\pi(p) = \emptyset$, necessarily $\tilde z = 0$, and we get from (6) that $z(p,v) = -S_\pi^\dagger(p) g_\pi^v(p)$. If $W_k = w_k I$, i.e. the weights are uniform and no data is missing, $S_i = w_i I / \|p - a_i\|$, and this step reduces to the the conventional Weiszfeld step used in (2).

ii) When $\pi(p) = \{k\}$ is a singleton, a $\tilde z$ may be easily found by minimising $h(z, v; p) = g_\pi^v(p)^T z + \|z\|_k$ over $\{z \in \mathcal{R}(W_k) \mid \|z\|_k = 1\}$. By positive homogeneity of h, its minimum value is zero over this set exactly when it is the same over $Z(p)$, so that we may choose $\tilde z = 0$ in this case. The result is therefore $\tilde z = -(W_k^\dagger)^2 g_\pi^v(p) / \left\| W_k^\dagger g_\pi^v(p) \right\| \in Z(p)$ when $\left\| W_k^\dagger g_\pi^v(p) \right\| \geq 1$, and $\tilde z = 0$ otherwise.

iii) When $\#\pi(p) > 1$, but the data do not overlap, i.e. $\mathcal{R}(W_i) \cap \mathcal{R}(W_j) = \{0\}$ for distinct $i, j \in \pi(p)$, $\tilde z$ can be calculated independently on each $\mathcal{R}(W_i)$, with the above result. This case is of importance in our application examples, and also in relation to the convergence results below.

iv) When $\#\pi(p) > 1$, but the data overlaps, the determination of appropriate $\tilde z$ is more complicated. However, in practical data sets, it is rare to have multiple vertices with partial coinciding information, furthermore agreeing with the current iterate. Appendix B in any case establishes relevant formulae for the non-partially-overlapping/hierarchical case.

2.2 Optimality conditions and the method

Let us now examine when $0 \in \partial^\circ f_\nu(p)$, where ∂° denotes the Clarke subdifferential; cf. [8, 20]. A necessary condition for this is that $0 \in \partial f(p) - \partial\nu(p) \triangleq \{g - v \mid g \in \partial f(p), v \in \partial\nu(p)\}$, or, equivalently, $\partial f(p) \cap \partial\nu(p) \neq \emptyset$. This is because $\partial^\circ(f-\nu)(p) \subset \partial^\circ f(p) + \partial^\circ(-\nu)(p)$ and $\partial^\circ(-\nu)(p) = -\partial\nu(p)$. But this means: for some $v \in \partial\nu(p)$, the convex function $f - v^T : p \mapsto f(p) - v^T p$ (and then $f_{\tilde{\nu}_p^v}$) has minimum at p. Therefore we define:

Definition 1. The point p is *semi-critical* for f_ν if $\partial f(p) \cap \partial\nu(p) \neq \emptyset$. The set of semi-critical points for our problem of interest is denoted P_∂.

By Theorem 1, p is then semi-critical if and only if $h(z, v; p) \geq 0$ for all $z \in \mathbb{R}^m \setminus \{0\}$ for some (fixed) $v \in \partial\nu(p)$.

Semi-criticality isn't generally sufficient for criticality in the sense $0 \in \partial^\circ f_\nu(p)$ (let alone local minimality). However, whenever either f or ν is differentiable at p, then by [8, Proposition 2.3.3], convexity and finiteness, $\partial^\circ f_\nu(p) = \partial f(p) - \partial\nu(p)$, and thus semi-criticality is equivalent to criticality. In particular, this holds whenever $\pi(p) = \emptyset$. On the other hand, if some upper convexification of f_ν by $\tilde{\nu}_p^v$ does not have a minimum at p, it then has a direction of descent, and so has f_ν. We can improve from semi-criticality a bit, however. Recall that a set-valued mapping F is outer-semicontinuous [24], if $p_i \to p$ and $v_i \in F(p_i)$, imply that every accumulation point of $\{v_i\}$ is in $F(p)$.

Definition 2. Let $\mathcal{D}\nu$ be an outer-semicontinuous mapping, such that $\emptyset \neq \mathcal{D}\nu(p) \subset \partial\nu(p)$, for $p \in \mathbb{R}^m$. If $\partial f(p) \cap \mathcal{D}\nu(p) \neq \emptyset$, we refer to p as *$\mathcal{D}$-critical* for f_ν. The set of $\mathcal{D}$-critical points for our problem of interest is denoted $P_\mathcal{D}$.

By Theorem 1, $\mathcal{D}$-criticality is equivalent to $h(z, v; p) \geq 0$ holding for all z for some $v \in \mathcal{D}\nu(p)$. The maximal system of such sets is, of course, the system $\partial\nu$ (as the subdifferential of a finite convex function is outer-semicontinuous). The minimal system is of necessity

$$\mathcal{D}_N\nu(p) \triangleq \{\lim_{i\to\infty} \nabla\nu(p_i) \mid p_i \to p, \nu \text{ is differentiable at } p_i\},$$

the convex hull of which is $\partial\nu(p)$.

These considerations finally lead us to extend the SOR-Weiszfeld iteration for incomplete data as follows.

Algorithm 1 (The perturbed SOR-Weiszfeld method)**.**

1. Set $r = 0$, and choose an initial iterate $p_0 \in \mathbb{R}^m$. Choose $\mathcal{D}\nu$ satisfying Definition 2 (typically $\partial\nu$ or $\mathcal{D}_N\nu$), as well as a stopping criterion.
2. Choose $v_r \in \mathcal{D}\nu(p_r)$, $\tilde{z} \in Z(p_r)$, $\omega \in (1, 2)$ and $\alpha \in (0, \alpha_0(\omega, \tilde{z}, v_r; p_r))$, as described in Lemma 1.
3. Calculate $p_{r+1} \triangleq T_\omega(p_r, v_r)$ with z defined by (6), and

$$T_\omega(p, v) \triangleq p + \omega z(p, v).$$

4. If the stopping criterion is not satisfied, continue from Step 2 with $r \triangleq r + 1$.

The choice of $v_r \in \mathcal{D}\nu(p_r)$ is arbitrary because we only have partial convergence to $\mathcal{D}$-critical points, and if there is a single v_r for which $f_{\tilde{\nu}_{p_r}^{v_r}}$ has no direction of descent, we have found such a point.

Lemma 2. *The iteration T_ω is descending for f_ν if $p \notin P_\mathcal{D}$.*

Proof. By Lemma 1, $z(p, v)$ is a direction of descend for $f_{\tilde{\nu}_p^v}$ when $v \notin \partial f(p)$, and therefore for f_ν as well. □

2.3 Convergence

We now turn to the convergence properties. The following Lemma 3 is an essential part that tells us that the iterates deflect from clusters of vertices at distance from $P_{\mathcal{D}}$. This along with some additional assumptions on choice of step length and the form of f_ν, allows us to exploit the continuity of T_ω on a subspace to show the convergence to $\mathcal{D}$-critical points in Theorem 2, assuming the iterates do not diverge.

We denote $p' \triangleq T_\omega(p, v)$. We will sometimes omit v from the parameters for brevity, and write $\tilde{z}(p)$ etc. The specific selection is denoted $v(p)$. The closed ball of radius δ centred at $x \in \mathbb{R}^m$, is denoted by $B(x, \delta)$.

Lemma 3. *Let the points and subgradients $p_r \in \mathbb{R}^m, v_r \in \mathcal{D}\nu(p_r)$ $(r = 1, 2, \ldots)$ and $q \in \mathbb{R}^m, u \in \mathcal{D}\nu(q)$ be given, with constant $\pi' \triangleq \pi(p_r) \subsetneq \pi(q)$. Suppose that $\tilde{z} \in Z(q)$ with (i) $\rho_{\pi'}\tilde{z} = 0$, and (ii) $h(\tilde{z}, u; q) < 0$. If (p_r, v_r) converge to (q, u), then for all $\omega \geq 1$ and some $k \in \pi_{\tilde{z}} \triangleq \{k \in \pi(q) \setminus \pi' \mid W_k\tilde{z} \neq 0\}$, it holds that*

$$\limsup_{r\to\infty} \frac{d_k(p'_r)}{d_k(p_r)} > 1. \tag{9}$$

In fact, $\liminf_{r\to\infty} \sup_{k\in\pi_{\tilde{z}}} d_k(p'_r)/d_k(p_r) > 1$, since we may apply the argument to any subsequence of the original.

Proof. Denote $p = p_r$ and $v = v_r$ for arbitrary r, for lighter notation. We may write

$$g_\pi(p) = \sum_{i\notin\pi'} S_i(p)(p - a_i) = \left(\sum_{i\notin\pi'} S_i(p)(q - a_i) + S_\pi(p)(p - q)\right).$$

Since $\bar{\rho}_{\pi'} S_\pi^\dagger(p) S_\pi(p) = \bar{\rho}_{\pi'}$ by our prevailing assumption $\sum \mathcal{R}(W_k) = \mathbb{R}^m$, as well as $\rho_{\pi'}(p - q) = 0$, we have according to (6) that

$$\bar{\rho}_{\pi'}(p' - q) = p - q - \omega\bar{\rho}_{\pi'} S_\pi^\dagger(p) g_\pi^v(p) = (1 - \omega)(p - q) - \omega\bar{\rho}_{\pi'} S_\pi^\dagger(p)\tilde{g}^{v(p)}(p), \tag{10}$$

where $\tilde{g}^v(p) \triangleq \sum_{i\notin\pi(q)} S_i(p)(q - a_i) - v$.

Let now $k \in \pi_{\tilde{z}}$. Since $W_k q = W_k a_k$, (9) follows if

$$\limsup_{r\to\infty} \frac{\|\bar{\rho}_{\pi'}(p'_r - q)\|_k}{\|p_r - q\|_k} > 1.$$

Thus, by applying $\omega \geq 1$ and the reverse triangle inequality to the W_k-norm of (10), it becomes sufficient to show that for some k, $\limsup_{r\to\infty}(\omega N_k(p_r) - |1 - \omega|) > 1$, i.e. $\limsup_{r\to\infty} N_k(p_r) > 1$, where

$$N_k(p) \triangleq \left\| \bar{\rho}_{\pi'} \left(\sum_{i\notin\pi'} W_i^2 \frac{d_k(p)}{d_i(p)}\right)^\dagger \tilde{g}^{v(p)}(p) \right\|_k.$$

Suppose $\limsup_r N_k(p_r) \leq 1$ for all $k \in \pi_{\tilde{z}}$, and choose $\epsilon > 0$. Then, for sufficiently large r, since $\|\tilde{z}\|_k = 0$ for $k \in (\pi(q) \setminus \pi') \setminus \pi_{\tilde{z}}$, an application of the

Cauchy-Schwarz inequality shows

$$
\begin{aligned}
\epsilon + \sum_{k\in\pi(q)\setminus\pi'} \|\tilde{z}\|_k &\geq \sum_{k\in\pi(q)\setminus\pi'} \|\tilde{z}\|_k N_k(p_r) \\
&\geq - \sum_{k\in\pi(q)\setminus\pi'} \tilde{z}^T W_k^2 (\frac{\bar{\rho}_{\pi'}}{d_k(p_r)}) \left(\sum_{i\notin\pi'} \frac{\bar{\rho}_{\pi'} W_i^2}{d_i(p_r)} \right)^{\dagger} \tilde{g}^{v_r}(p_r) \qquad (11) \\
&= -\tilde{z}^T \left(\sum_{k\in\pi(q)\setminus\pi'} \Gamma_k(p_r) \right) \left(\sum_{i\notin\pi'} \Gamma_i(p_r) \right)^{\dagger} \tilde{g}^{v_r}(p_r),
\end{aligned}
$$

where $\Gamma_i(p) \triangleq W_i^2 \bar{\rho}_{\pi'} x(p)/d_i(p)$, and $x(p) \triangleq 1/\left\| \sum_{k\in\pi(q)\setminus\pi'} W_k^2 \bar{\rho}_{\pi'}/d_k(p) \right\|$ is a normalising factor.

Observe that $\Gamma_i(p_r) \to 0$ for $i \notin \pi(q)$, faster than for $i \in \pi(q) \setminus \pi'$ (if such were to happen). Therefore $\sum_{i\notin\pi'} \Gamma_i(p_r) - \sum_{k\in\pi(q)\setminus\pi'} \Gamma_k(p_r) \to 0$, and likewise for the pseudo-inverses. Now letting $\epsilon \searrow 0$ and going to the limit in (11) yields

$$
\sum_{k\in\pi(q)\setminus\pi'} \|\tilde{z}\|_k \geq -\tilde{z}^T \bar{\rho}_{\pi'} \rho_{\pi(q)} g_\pi^u(q).
$$

This combined with assumption (i) says that $h(\tilde{z}, u; q) \geq 0$, in contradiction to assumption (ii). □

Lemma 4. *Suppose* $(p_r, v_r) \to (q, u)$ *with constant* $\pi(p_r) = \pi'$ *and* $\tilde{z}(p_r) = 0$. *Then we may take* $\rho_{\pi'} \tilde{z}(q) = 0$.

Proof. Since $\rho_{\pi'}(q - p_r) = 0$, we have as $r \to \infty$ that

$$
\rho_{\pi'} g_\pi(p_r) = \rho_{\pi'} \sum_{i\notin\pi'} S_i(p_r)(p_r - a_i) = \rho_{\pi'} \sum_{i\notin\pi(q)} S_i(p_r)(p_r - a_i) \to \rho_{\pi'} g_\pi(q).
$$

Consequently, for $\tilde{z} \in Z(q)$,

$$
g_\pi^u(q)^T \rho_{\pi'} \tilde{z} + \sum_{k\in\pi'} \|\tilde{z}\|_k = \lim_{r\to\infty} g_\pi^v(p_r)^T \rho_{\pi'} \tilde{z} + \sum_{k\in\pi'} \|\tilde{z}\|_k = \lim_{r\to\infty} h(\tilde{z}, v_r; p_r) \geq 0,
$$

with the inequality holding by $\tilde{z}(p_r) = 0$. Therefore, we can take $\rho_{\pi'} \tilde{z} = 0$, as any other choice would increase the value of the remaining $\|\tilde{z}\|_k$ for $k \in \pi(q) \setminus \pi'$ in $h(\cdot, u; q)$. □

Assumption 1. The set of iterates $\{p_r \mid r = 1, 2, \ldots\}$ generated by Algorithm 1 is bounded. The function f_ν is bounded from below. The step sizes ω_r satisfy the conditions of Algorithm 1, and there exists $\overline{\omega} < 2$, such that $\omega_r \in [1, \overline{\omega}]$. Furthermore, $\tilde{z}(p_r) = 0$ (i.e. $\pi(p_r') \supset \pi(p_r)$) eventually.

Lemma 5. *The step sizes can be chosen to satisfy* $\tilde{z}(p_r) = 0$ *eventually. Hence eventually* $\pi(p_r) = \pi'$ *is constant.*

Proof. Choose ω (eventually) so as to avoid adding elements to $\pi(p_r)$. This can be done, since in each direction $z(p_r, v_r)$, there are finitely many step lengths for which $d_k(p_r) = 0$ for some $k \notin \pi(p_r)$. Then $\pi(p_r') \subset \pi(p_r)$, which can be strict only finitely many times, exactly when $\tilde{z}(p_r) \neq 0$. □

Lemma 6. *Suppose Assumption 1 holds, and let* (q, u) *be a cluster point of* $\{(p_r, v_r)\}$*. Then* $q \in P_{\mathcal{D}}$*, if* $h(\cdot, u; q) \geq 0$ *on* $Z(q)$*.*

Proof. Since $\{f_\nu(p_r)\}$ is bounded from below by assumption, and monotonically decreasing by Lemma 2, it holds that

$$\lim_{r\to\infty}(f_\nu(p_r) - f_\nu(p'_r)) = 0. \tag{12}$$

Let then $\{(p_{r_\ell}, v_{r_\ell})\}$ be a subsequence convergent to (q, u). If $p_{r_\ell} \in P_{\mathcal{D}}$ for some ℓ, then there's nothing to prove, so suppose this is not the case. We may assume that $\pi(p_{r_\ell}) = \pi'$ is constant. Also, since for $i \notin \pi(q)$ it holds that $d_i(q) > 0$, we must have $\pi(p_{r_\ell}) \subset \pi(q)$, whence $\rho_{\pi'}(q - p_{r_\ell}) = 0$.

If $\pi(q) = \emptyset$, then also $\pi' = \emptyset$. If q were not $\mathcal{D}$-critical, it would hold that $f_\nu(q') < f_\nu(q)$ for all choices of $v(q) \in \mathcal{D}\nu(q)$ and $\omega \in [1, \overline{\omega}]$. But since T_ω for fixed ω is continuous around (q, u), and since $(p_{r_\ell}, v_{r_\ell}) \to (q, u)$, we get $T_\omega(p_{r_\ell}, v_{r_\ell}) \to T_\omega(q, u)$. Therefore, $\lim f_\nu(p'_{r_\ell}) = f_\nu(q') < f_\nu(q) = \lim f_\nu(p_{r_\ell})$, which contradicts (12). Thus $T_\omega(q, u) = q$, and consequently in the case of varying $\omega_r \in [1, \overline{\omega}]$, we see that the line segment $[T_1(p_{r_\ell}, v_{r_\ell}), T_{\overline{\omega}}(p_{r_\ell}, v_{r_\ell})] \ni T_{\omega_{r_\ell}}(p_{r_\ell}, v_{r_\ell}) = p'_{r_\ell}$ vanishes at the limit. Therefore, $q \in P_{\mathcal{D}}$.

Suppose then that $\pi(q) \neq \emptyset$. Since $h(\cdot, u; q) \geq 0$ over $Z(q)$, we have $\tilde{z}(q, u) = 0$, and it remains to show that $z(q, u) = 0$, i.e. $\bar{\rho}_{\pi(q)} S_\pi^\dagger(q) g_\pi^u(q) = 0$. We have $\rho_{\pi(q)} \bar{\rho}_{\pi'} z(p_{r_\ell}) = -\rho_{\pi(q)} \bar{\rho}_{\pi'} S_\pi^\dagger(p_{r_\ell}) g_\pi^{v_{r_\ell}}(p_{r_\ell}) \to 0$, because $v_{r_\ell} \to u$ is bounded, and $\rho_{\pi(q)} \bar{\rho}_{\pi'} S_\pi^\dagger(p_{r_\ell})$ goes to zero (with $1/d_k(p_{r_\ell})$ going to infinity in $S_\pi(p_{r_\ell})$ for $k \in \pi(q) \setminus \pi'$). As $\bar{\rho}_{\pi(q)} S_\pi^\dagger(p_{r_\ell}) g_\pi^{v_{r_\ell}}(p_{r_\ell})$ does not depend on a_k for $k \in \pi(q)$, it is convergent. Therefore, in summary, we have $\bar{\rho}_{\pi'} S_\pi^\dagger(p_{r_\ell}) g_\pi^{v_{r_\ell}}(p_{r_\ell}) \to \bar{\rho}_{\pi(q)} S_\pi^\dagger(q) g_\pi^u(q)$.

Consequently, $\lim_\ell T_\omega(p_{r_\ell}, v_{r_\ell}) = T_\omega(q, u)$ for fixed ω, the choice of α being irrelevant because $\tilde{z}(p_{r_\ell}, v_{r_\ell}) = \tilde{z}(q, u) = 0$.[1] Now the same argument as was used in the case $\pi(q) = \emptyset$ applies. We therefore have $q \in P_{\mathcal{D}}$. □

Lemma 7. *Suppose Assumption 1 holds. Let* Q_k *denote the set of cluster points* (q, u) *of* $\{(p_r, v_r)\}$*, such that* $k \in \pi(q) \setminus \pi'$*. We have,*

i) If $\liminf_{\ell\to\infty} d_k(p'_{r_\ell})/d_k(p_{r_\ell}) > 1$ *for all subsequences approaching* Q_k*, then* $Q_k = \emptyset$*.*

ii) The above condition follows if for each $(q, u) \in Q_k$*, there exists* $\tilde{z} \in Z_k(q) \triangleq \{\tilde{z} \in Z(q) \mid W_i \tilde{z} = 0 \text{ for } i \in \pi(q) \setminus \{k\}\}$ *such that* $h(\tilde{z}, u; q) < 0$*.*

Proof. Note that Q_k is compact by boundedness of $\{(p_r, v_r)\}$, and that $Z_k(q) = \{\tilde{z} \in Z(q) \mid \rho_{\pi'} \tilde{z} = 0, \pi_{\tilde{z}} = \{k\}\}$. Let $\{(p_{r_\ell}, v_{r_\ell})\}$ be a subsequence of $\{(p_r, v_r)\}$ approaching Q_k (with constant $\pi(p_{r_\ell}) = \pi'$). Under the conditions of (ii), we must have $\liminf_{\ell\to\infty} d_k(p'_{r_\ell})/d_k(p_{r_\ell}) > 1$, because otherwise we could find a subsequence convergent to some $(q, u) \in Q_k$, for which an application of Lemma 3 would yield $h(\tilde{z}, u; q) \geq 0$ for all $\tilde{z} \in Z_k(q)$, in contradiction to our assumptions.

We may therefore assume that there exist $\delta > 0$ and $\epsilon > 0$, such that whenever $(p_r, v_r) \in Q_k + B(0, \delta)$, then $d_k(p'_r) \geq (1 + \epsilon) d_k(p_r)$. Therefore, since $d_k(p_r) > 0$, there exists a $t > r$ such that $(p_{[t]}, v_{[t]}) \notin Q_k + B(0, \delta)$. Thus the whole sequence cannot converge to Q_k.

[1] In this lemma, $\alpha \searrow 0$ would suffice, instead of $\tilde{z} = 0$. This could be explicitly assumed, but also follows from convergence assumptions, and sometimes from (7). The argument of Lemma 7 could also be extended to allow $k \in \pi(p_r)$, provided $\|\alpha \tilde{z}\|_k > 0$ for a subsequence. However, application/variant of Lemma 4 would demand additional assumptions.

There then exists a subsequence $\{(p_{r_\ell}, v_{r_\ell})\}$ with $(p_{r_\ell}, v_{r_\ell}) \notin Q_k + B(0,\delta)$, and $(p'_{r_\ell}, v'_{r_\ell}) \in Q_k + B(0,\delta)$. Since Q_k contains all the cluster points with $k \in \pi(q)$, there also exists $\delta' > 0$ such that $d_k(p_{r_\ell}) > \delta'$. Therefore, if there's a subsequence convergent to Q_k, we must have $d_k(p'_{r_\ell}) \to 0$. But, since the algorithm moves from p_r to a direction of descent of $f_{\tilde{\nu}^{v_r}_{p_r}}$, we have

$$f_\nu(p_r) - f_\nu(p'_r) \geq f_{\tilde{\nu}^{v_r}_{p_r}}(p_r) - f_{\tilde{\nu}^{v_r}_{p_r}}(p'_r) = \frac{1}{2} \sum_{i \notin \pi(p_r)} (d_i(p_r) - d_i(p'_r))^2 / d_i(p_r), \quad (13)$$

where the final estimate and term $C/2$ are from the proof of Theorem 1. This with $r = r_\ell$ provides a contradiction to (12). Therefore $Q_k = \emptyset$. □

Theorem 2. *Suppose Assumption 1 holds, and that for all $\pi \in \mathcal{R}(\pi(\cdot))$, $k, i \in \pi$, $k \neq i$ implies $\mathcal{R}(W_k) \cap \mathcal{R}(W_i) = \{0\}$. Then either $\{(p_r, v_r)\}$ has a cluster point (q, u) with $q \in P_{\mathcal{D}}$, or the sequence diverges.*

Proof. If there exists a cluster point (q, u), such that $h(\cdot, u; q) \geq 0$ on $Z(q)$, Lemma 6 proves the claim.

Otherwise, to reach a contradiction, we may assume that $(p_r, v_r) \to (q, u)$, where $h(\tilde{z}, u; q) < 0$ for some $\tilde{z} \in Z(q)$. According to Lemma 4, we may take $\rho_{\pi'}\tilde{z}(q) = 0$. Furthermore, on the assumption that $\mathcal{R}(W_k) \cap \mathcal{R}(W_i) = \{0\}$ for $k, i \in \pi(q)$, $h(\cdot, u; q)$ is independent on each $\mathcal{R}(W_k)$. We may therefore choose $\tilde{z} \in Z_k(q)$ for some $k \in \pi(q) \setminus \pi'$. An application of Lemma 7 to $Q_k = \{(q, u)\}$ now provides the desired contradiction. □

Remark 1. We have the following further observations:

i) If $p_r \to q$, but $\{v_r\}$ diverges, then ν must be nondifferentiable at q.
ii) If a cluster point has $\pi(q) = \pi'$, $q \in P_{\mathcal{D}}$ (by Lemma 6). In particular, any cluster point with $\pi(q) = \emptyset$, is a solution.
iii) If $\#\pi(p) \leq 1$ for all $p \in \mathbb{R}^m$, then there is a cluster point $q \in P_{\mathcal{D}}$. (Combine Lemmas 6 and 7.) This is unfortunately not the case in our forthcoming applications with "lifted" data.
iv) If there are multiple cluster points with differing $\pi(q)$, there are actually infinitely many of them: for some k, there are iterates with both $d_k(p_r) > \delta$, as well as $d_k(p'_{r_\ell}) \in [\delta/2, \delta)$, since $d_k(p'_{r_\ell}) \to 0$ is not possible by (13). Therefore there are cluster points in this distance range. Now let $\delta \searrow 0$.

2.4 Boundedness

For the above partial convergence results to be of any use, an easily checkable condition is needed to ensure that f_ν is bounded from below, and that there are cluster points: the iterates stay bounded. Because the sequence $\{f_\nu(p_r)\}_{r=1}^\infty$ is descending, it suffices to show that the level sets of f_ν are bounded. This is where we need the general results of Appendix A, relating $\operatorname{cl}\mathcal{R}(\partial\nu) \subset \operatorname{int}\mathcal{R}(\partial f)$ to this. To apply these results, we need to calculate the boundary of $\mathcal{R}(\partial f)$ for f defined by (3). We denote by $\operatorname{cl} A$, $\operatorname{bd} A$, and $\operatorname{int} A$, the closure, border, interior of the set A, respectively.

Lemma 8. *Let $A \triangleq \bigcup_{p \in \mathbb{R}^m} \partial f(p)$. Then $\operatorname{cl} A$ is convex and bounded, and*

$$\operatorname{bd} A = Z \triangleq \bigcup_{\pi_b} Z_{\pi_b},$$

with the union taken over $\pi_b \subset \{1, \ldots, n\}$ *such that* $\mathcal{R}(\rho_{\pi_b}) \subsetneq \mathbb{R}^m$ *and* $k \in \pi_b$ *whenever* $\mathcal{R}(W_k) \subset \mathcal{R}(\rho_{\pi_b})$*. Here*

$$Z_{\pi_b} = \{ \sum_{k \notin \pi_b} W_k^2 q / \|q\|_k + v \mid q \in Q_{\pi_b}, v \in \operatorname{cl} A_{\pi_b} \},$$

$$Q_{\pi_b} \triangleq \{q \in \mathbb{R}^m \mid W_j q = 0 \ (j \in \pi_b), W_k q \neq 0 \ (k \notin \pi_b)\},$$

$$A_{\pi_b} \triangleq \bigcup_{p \in \mathbb{R}^m} \partial(\sum_{k \in \pi_b} d_k)(p).$$

Proof. The subdifferentials of f are clearly uniformly bounded: for $g \in \partial f(p)$, $\|g\| \leq \sum_{k=1}^n \|W_k\|$. Hence A is bounded. By e.g. [23, Section 24] $\operatorname{cl} A$ is also convex. It remains to prove that $\operatorname{bd} A$ is of the claimed form.

Let $q \neq 0$. Then $\max_{g \in \operatorname{cl} A} g^T q$ is attained by any $g \in Z_{\pi_b}$ obtained (as seen by considering $q^T \nabla(q^T \nabla f)(p) = q^T \nabla^2 f(p) q)$ as a limit of some sequence $g_i \in \partial f(p_i)$ as $\|p_i\| \to \infty$ with $W_k^2(p_i - a_k)/\|p_i - a_k\|_k \to W_k^2 q/\|q\|_k$, when $W_k q \neq 0$. It then has the form

$$g = \sum_{k \notin \pi_b} W_k^2 q / \|q\|_k + v \tag{14}$$

with $\pi_b = \{j \in \{1, \ldots, n\} \mid W_j q = 0\}$, and $v \in \operatorname{cl} A_{\pi_b}$. Therefore, all the exposed faces of $\operatorname{cl} A$ are contained in the sets Z_{π_b}, that are closures of unions of these faces. It remains to prove that their union forms all of $\operatorname{bd} A$.

The exposed faces of $\operatorname{cl} A$ are precisely the sets of the form $\operatorname{cl} A \cap H$, where H is a supporting hyperplane to $\operatorname{cl} A$; see [23]. But $\operatorname{cl} A$ is the intersection of the corresponding half-spaces. Thus, if $g \in \operatorname{cl} A$ has a ball $B(g, \epsilon)$ around it that is not intersected by any of the hyperplanes H (and thus not by any of the Z_{π_b}), then $g \notin \operatorname{bd} A$. Otherwise, since the intersecting hyperplanes are defined by a compact set of parameters (closed subset of $\operatorname{bd} A \times \operatorname{bd} B(0,1)$), we may find a supporting hyperplane H that contains g. But then $g \in \operatorname{cl} A \cap H$, an exposed face. □

3 Applications: bi-objective clustering

Consider a multiobjective formulation of the multifacility location problem:

$$\min_{\bar{p} \in (\mathbb{R}^m)^s} (f_1, f_2)(\bar{p}; \bar{a}), \tag{15}$$

where the minimum is in the sense of Pareto-optimality, $\bar{p} = (p_1, \ldots, p_s) \in (\mathbb{R}^m)^s$, and $\bar{a} = (a_1, \ldots, a_n) \in (\mathbb{R}^m)^n$. The objectives are defined as

$$f_1(\bar{p}) = \sum_{i=1}^{s} \sum_{j=1}^{n} d_j(p_i), \qquad f_2(\bar{p}) = -\sum_{i=1}^{s} \sum_{j=1}^{s} d(p_j, p_i)$$

for some distance functions d_j dependent on a_j, and d. The objective f_1 indicates our desire to place cluster centres $\{p_j\}$ as close to the data as possible as defined by means of the distances d_j, while f_2 indicates our desire to place the cluster centres as far apart from each other as possible. (We want to minimise f_1 and at the same time maximise $-f_2$.)

3.1 Squared Euclidean distance

Although it does not fit in the framework derived in the earlier sections, for comparison to what will follow and also to the classical K-means (which uses the same distance; see [10]), we will first consider the case when $d(x,y) = \frac{1}{2}\|x-y\|^2$ is the squared distance. For simplicity we limit ourselves to the case of complete information, $d_j = d(a_j, \cdot)$. We then get as the Karush-Kuhn-Tucker necessary condition for Pareto optimality (see e.g. [19, Chapter I.3]) that

$$\lambda_1 \sum_{j=1}^{n} (p_i - a_j) - \lambda_2 \sum_{j=1}^{s} (p_i - p_j) = 0 \quad \text{for all } i = 1, \ldots, s,$$

or that

$$(\lambda_1 n - \lambda_2 s) p_i - \lambda_1 \sum_{j=1}^{n} a_j + \lambda_2 \sum_{j=1}^{s} p_j = 0 \quad \text{for all } i = 1, \ldots, s, \tag{16}$$

for some $\lambda_1, \lambda_2 \geq 0$ with strict inequality for at least one of λ_1 or λ_2.

If $\lambda_1 n = \lambda_2 s$, we get the solution candidates

$$\frac{1}{s} \sum_{i=1}^{s} p_i = \frac{1}{n} \sum_{j=1}^{n} a_j. \tag{17}$$

If, on the other hand $\lambda_1 n - \lambda_2 s \neq 0$, we find that all the $\{p_i\}_{i=1}^{s}$ are equal by subtracting the term (16) for p_i and p_j $(i \neq j)$. Hence, unless $\lambda_1 = 0$, in fact (17) holds. In the case $\lambda_1 = 0$ there is no finite minimum, so we may ignore it.

Let us now check when solutions of (17) are Pareto-optimal. Expand the expressions for d to yield

$$f_1(\bar{p}) = \sum_{i=1}^{s} \sum_{j=1}^{n} \frac{1}{2} \big(\|p_i\|^2 + \|a_j\|^2 \big) - \big(\sum_{i=1}^{s} p_i \big)^T \big(\sum_{j=1}^{n} a_j \big)$$

and

$$f_2(\bar{p}) = -\sum_{i=1}^{s} \sum_{j=1}^{s} \frac{1}{2} \big(\|p_i\|^2 + \|p_j\|^2 \big) + \big(\sum_{i=1}^{s} p_i \big)^T \big(\sum_{j=1}^{s} p_j \big).$$

Thus if (17) holds, then both f_1 and f_2 have a constant term at the end and f_2 decreases iff $\sum_i \|p_i\|^2$ increases. But since this means that f_1 increases, the solutions of (17) are precisely the Pareto-optimal solutions of the original problem. This says that the Pareto-optima are where the cluster centre means equal the data means. The condition for Pareto-optimality is therefore very weak, and the solutions are abundant.

3.2 Euclidean distance

With the Euclidean distance $d(x,y) = \|x-y\|$, and d_j defined as in Section 2, we get more interesting results. The scalarisation of the problem (15) by the factor $\lambda \geq 0$ then reads as

$$\min f_1(\bar{p}) + \lambda f_2(\bar{p}) \tag{18}$$

(cf. [19, Section II.3.1]). This problem can be cast as a problem of finding a perturbed spatial median as follows. For each $i = 1, \ldots, s$ and $j = 1, \ldots, n$, let

$$a_j^i \triangleq (\underbrace{0, \ldots, 0}_{m(i-1) \text{ times}}, a_j^T, \underbrace{0, \ldots, 0}_{m(s-i) \text{ times}})^T,$$

and W_j^i be such that $W_j^i(\bar{p} - a_j^i) = W_j(p_i - a_j)$. Then

$$f_1(\bar{p}) = \sum_{i,j} \left\| W_j^i(\bar{p} - a_j^i) \right\| .$$

Because f_2 is concave and finite, the problem can be modelled as a perturbed spatial median problem with vertices $\{a_j^i\}$ and perturbation $-\lambda f_2$. Note that if $\{W_j\}$ satisfy the range non-overlap assumption of Theorem 2, so do $\{W_j^i\}$. Hence, by Theorem 2, Theorem 7, and Lemma 11 in Appendix A, Algorithm 1 is applicable for finding semi-critical points (Kuhn-Tucker points of the multiobjective problem), if we can bound $\mathcal{R}(\partial(-\lambda f_2))$ within $\mathcal{R}(\partial f_1)$.

With f denoting here and throughout the paper, the function defined by (3) with the original data $\{a_j\}$, not $\{a_j^i\}$, we note that $\partial f_1(\bar{p}) = \partial f(p_1) \times \cdots \times \partial f(p_s)$, since f_1 consists of s sets of n terms depending on different components of $\bar{p}$. Also, since f_2 is positively homogeneous, we have $\mathcal{R}(\partial(-\lambda f_2)) = \partial(-\lambda f_2)(0)$. Now, since $\mathcal{R}(\partial f_1)$ is a product space, it suffices to consider the slices $[\partial(-\lambda f_2)(0)]_i$ of this subdifferential independently. At differentiable points

$$[\nabla(-\lambda f_2)(\bar{p})]_i = 2\lambda \sum_{j \neq i} \frac{p_i - p_j}{\|p_i - p_j\|}.$$

Therefore, by the limit characterisation of the subdifferential, it suffices to check that

$$\lim_{q_1, \ldots q_{s-1} \to 0} 2\lambda \sum_{j=1}^{s-1} \frac{q_j}{\|q_j\|} \in \operatorname{int} \mathcal{R}(\partial f)$$

or that

$$B(0, 2\lambda(s-1)) \in \operatorname{int} \mathcal{R}(\partial f).$$

In the simple case with $W_k = I$ for all $k = 1, \ldots, n$, this follows if $2\lambda < n/(s-1)$ (when $s > 1$), because $\operatorname{cl} \mathcal{R}(\partial f) = B(0, n)$ then. For incomplete and weighted data, we must consider the "minimal dimension" of A: by Lemma 8, we must find minimum $\|z\|$ for $z = \sum_{k \notin \pi_b} W_k^2 q / \|q\|_k + v \in Z_{\pi_b}$, among all π_b. The sets Q_{π_b} and A_{π_b} are orthogonal, and the v can be made arbitrarily close to zero, being a subgradient of a reduced spatial median problem. Therefore, it can and must be chosen to be zero, and the remaining sum sets the bound. Thus we may state:

Theorem 3.

i) The level sets of the scalarised problem (18) *are bounded if* $0 \leq 2\lambda < \beta/(s-1)$ *with* $\beta = \min \left\| \sum_{k \notin \pi_b} W_k^2 q / \|q\|_k \right\|$, *with the minimum taken over all* $q \in Q_{\pi_b}$ *and* $\pi_b \subset \{1, \ldots, n\}$ *satisfying the conditions of Lemma 8.*

ii) If, furthermore, $W_k = \rho_k$ *for zero-one diagonal matrices* ρ_k, $\beta \geq \min \#\pi_b^c$ *with* $\pi_b^c \triangleq \{1, \ldots, n\} \setminus \pi_b$. *In particular,* $\beta \geq \#\{\rho_k = I\}$.

Proof. Only the lower bound $\min \#\pi_b^c \leq \beta$ demands further proof. Since $\rho_{\pi_b} q = 0$, we have

$$\left\| \sum_{k\in\pi_b^c} W_k^2 q/\|q\|_k \right\| \geq \sqrt{\sum_{i:(\bar\rho_{\pi_b})_{ii}=1} q_i^2 \Big(\sum_{k\in\pi_b^c:(\rho_k)_{ii}=1} \frac{1}{\|\rho_k q\|}\Big)^2}$$
$$\geq \sqrt{\sum_{i:(\bar\rho_{\pi_b})_{ii}=1} q_i^2 \#\{k\in\pi_b^c : (\rho_k)_{ii}=1\}^2/\|q\|^2}$$
$$\geq \min_{i:(\bar\rho_{\pi_b})_{ii}=1} \#\{k\in\pi_b^c : (\rho_k)_{ii}=1\}.$$

If $(\rho_k)_{ii}=0$ and $(\bar\rho_{\pi_b})_{ii}=1$, then $(\bar\rho_{\pi_b\cup\{k\}})_{ii}=1$. Therefore, for some $\pi_{b'}\supset\pi_b\cup\{k\}$, with $\pi_{b'}\subsetneq\{1,\dots,n\}$ since $\rho_k\neq I$, both the set the minimum taken over is larger, as well as the values smaller. Therefore, taking the minimum over the admissible set of π_b as defined in Lemma 8, we get the first claimed lower bound. Finally, if $\mathcal{R}(\rho_k)$ is full, k is never contained in π_b. □

With such choices of λ as above, Algorithm 1 can thus in principle be applied to finding semi-critical points of the scalarised problem (18). We emphasise that Theorem 3(ii) provides a simple and explicit lower bound for the supremum of practical scalarisation values, as the amount of complete data. On the other hand, when $2\lambda>\beta/(s-1)$, $2\lambda\sum_{j=1}^{s-1} q_j/\|q_j\| \in \operatorname{cl}\mathcal{R}(\partial f)$ can be violated, whence $\mathcal{R}(\partial(-\lambda f_2))\not\subset\mathcal{R}(\partial f_1)$. Problem (18) is not bounded from below then, wherefore no finite pareto-optimal solution is generated by scalarisation parameters much larger than Algorithm 1 can be expected to handle.

Remark 2. Although we used the lifting of a_i to a_i^j in modelling the problem as a problem of perturbed spatial median, it is not necessary to work with such expanded data sets in practical implementations. Since the a_j^i for differing j have no coordinates with overlapping information, we have in particular that $g_\pi(\bar p) = (g_\pi(p_1),\dots,g_\pi(p_s))$ and $S_\pi(\bar p)=(S_\pi(p_1),\dots,S_\pi(p_s))$, where the right-hand-sides have been defined for the original data $\{a_i\}$. In consequence, there is no dependency between the p_j within the iterations of the SOR-Weiszfeld algorithm aside from calculating the "tilt" $v\in\partial(-f_2)(\bar p)$. Therefore, each iteration of Algorithm 1 can be calculated in parallel using the same step size for the different cluster centres after a subgradient of $-f_2$ has been calculated.

Remark 3. The convergent sequences of our method are to semi-critical points, not necessarily (local) minima. In addition to standard second degree conditions for a posteriori optimality checking, we do, however, have at least the following necessary optimality condition with a clear interpretation.

Lemma 9. *Suppose* $p_j=p_k$ $(j\neq k)$ *and* $\operatorname{rank}(\rho_\pi)<m$ *for* $\pi\triangleq\pi(p_j)=\pi(p_k)$. *Then* $\bar p$ *is not a local minimiser.*

Proof. The term $\|p_j-p_k\|$ is not differentiable at $p_j=p_k$, Therefore, with $\bar v=(v_1,\dots,v_s)$, there are multiple choices for v_j and v_k (dependent on each other) in all m dimensions, and we can in (5) choose v_j so that $[g_\pi(\bar p)]_j - v_j\neq 0$, and the same for k. Because $\operatorname{rank}(\rho_\pi)<m$, the term $\sum_{i\in\pi}\|z_j\|_i$ does not pose problems in forcing $h(\cdot;v,p)$ negative in (5). Thus the claim of the lemma follows from Theorem 1. □

4 Applications: the multisource Weber problem

The K-spatial median or the multisource Weber problem is a K-means type clustering criteria. Instead of the squared distance, the Euclidean distance is simply used. The standard formulation is

$$\min_{w_{ij},\bar{p}} \sum_{i=1}^{n} \sum_{j=1}^{s} w_{ij} d_i(p_j) \quad \text{with } w_{ij} \in \{0,1\} \text{ and } \sum_{j=1}^{s} w_{ij} = 1, \tag{19}$$

with K denoted by s here. The weights w_{ij} indicate to which cluster j the vertex i belongs to, and p_j is the cluster prototype.

The standard K-means-type algorithm [10, 25, 9] proceeds by assigning each a_i to the closest cluster centre p_j (setting $w_{ij} = 1$), calculating the spatial median p'_j for each of the clusters $A_j = \{a_i \mid w_{ij} = 1\}$, and repeating this until there is no change in the assignments. Convergence of this class of methods to (differentiable) Karush-Kuhn-Tucker points for some classes of distance functions in $\mathbb{R}^m$ is proved in [25], along with providing an extension to find local minima. The proof readily generalises to our case of incomplete data. For some other heuristic and local methods for solving the problem, see [9, 4, 2]. The global solution with outer approximation methods of the diff-convex formulation to be given below, is studied in [7]. Other approximation schemes are derived in [1].

Given the constraints on the weights, for fixed i, $\min_{w_{ij}} \sum_{j=1}^{s} w_{ij} d_i(p_j) = \min_{j=1,\ldots,s} d_i(p_j)$. Therefore, an alternative way to write (19) is

$$\min_{\bar{p}} \sum_{i=1}^{n} \min_{j=1,\ldots,s} d_i(p_j). \tag{20}$$

Because $\min\{x,y\} = x + y - \max\{x,y\}$, this formulation can be further recast as a DC problem by writing the objective function as

$$f_1(\bar{p}) + f_2^{\mathrm{KM}}(\bar{p}) \triangleq \big(\sum_{i=1}^{n} \sum_{j=1}^{s} d_i(p_j)\big) - \big(\sum_{i=1}^{n} \max_{j=1,\ldots,s} \sum_{k\neq j} d_i(p_k)\big).$$

But, indeed, using the lifting of a_i to a_i^j for $j = 1, \ldots, s$ as in Section 3.2, this problem is seen to be a problem of perturbed spatial median. This problem, however, has unbounded level sets: any change in p_j sufficiently far from the data when some other cluster centre is close to it, does not affect the function value. In other words, the problem may have "degenerate" solutions; cf. also [5]. Therefore, Theorem 7 (in Appendix A) can not be used to prove the applicability of our Weiszfeld-like algorithm. However, we can prove boundedness of the iterates directly with some conditions on the step sizes and the tilt $\bar{v}(\bar{p})$, after first analysing Algorithm 1 applied to this problem, in further detail.

4.1 Algorithm analysis and reduction

Let us calculate $\partial(-f_2)$. Similarly to the derivation of ∂f_1 in Section 3.2, we get

$$\partial\big(\sum_{k\neq j} d_i(p_k)\big)(\bar{p}) = \partial d_i(p_1) \times \cdots \partial d_i(p_{j-1}) \times \{0\} \times \partial d_i(p_{j+1}) \times \cdots \partial d_i(p_s)$$

and therefore, with $J_i \triangleq J_i(\bar{p})$ denoting the set of indices j for which $\sum_{k\neq j} d_i(p_k)$ reaches its maximum ($d_i(p_j)$ reaches minimum),

$$\begin{aligned}\partial(-f_2^{\mathrm{KM}})(\bar{p}) &= \bigcup_{\Lambda\in\mathcal{W}} \sum_{i=1}^{n} \sum_{j\in J_i} \lambda_{j,i}\partial\big(\sum_{k\neq j} d_i(p_k)\big)(\bar{p}) \\ &= \bigcup_{\Lambda\in\mathcal{W}} \sum_{i=1}^{n} \prod_{j=1}^{s} \left(\sum_{k\in J_i\setminus\{j\}} \lambda_{k,i} \right) \partial d_i(p_j) = \bigcup_{\Lambda\in\mathcal{W}} \sum_{i=1}^{n} \prod_{j=1}^{s} G_{j,i}\end{aligned} \tag{21}$$

with

$$G_{j,i} = \begin{cases} \partial d_i(p_j), & j \notin J_i, \\ (1-\lambda_{j,i})\partial d_i(p_j), & j \in J_i. \end{cases} \tag{22}$$

Here $\Lambda \triangleq \{\lambda_{j,i} \mid j \in J_i, i = 1,\dots,n\}$ and $\mathcal{W} \triangleq \mathcal{W}(\bar{p}) \triangleq \{\Lambda \mid \sum_{j\in J_i} \lambda_{j,i} = 1, \lambda_{j,i} \geq 0\}$. Also let $\mathcal{W}_{\mathrm{ext}} \triangleq \{\Lambda \mid \sum_{j\in J_i} \lambda_{j,i} = 1, \lambda_{j,i} \in \{0,1\}\}$ be the extreme points of $\mathcal{W}$.

After choosing the weights $\{\lambda_{j,i}\}$, we may therefore choose for $\bar{v}(\bar{p}) = \bar{v} = (v_1,\dots,v_s)$ each $v_j \in \sum_{i=1}^{n} G_{j,i}$ independently. Noting that $j \notin J_i$ implies $d_i(p_j) > 0$ and hence $i \notin \pi(p_j)$, let

$$v_j \triangleq \sum_{\substack{i\notin\pi(p_j) \\ J_i\ni j}} (1-\lambda_{j,i})\nabla d_i(p_j) + \sum_{\substack{i\notin\pi(p_j) \\ J_i\not\ni j}} \nabla d_i(p_j) + \sum_{i\in\pi(p_j)} (1-\lambda_{j,i}) W_i^2 z_j / \|z_j\|_i \,. \tag{23}$$

Then $\bar{v} \in \partial(-f_2^{\mathrm{KM}})(\bar{p})$, and $\bar{\rho}_{\pi(\bar{p})} g_\pi^v(\bar{p}) = (\bar{\rho}_{\pi(p_1)} g_1, \dots, \bar{\rho}_{\pi(p_s)} g_s)$ for

$$g_j = \sum_{i\notin\pi(p_j)} \nabla d_i(p_j) - \bar{\rho}_{\pi(p_j)} v_j = \sum_{\substack{i\notin\pi(p_j), \\ J_i\ni j}} \lambda_{j,i}\nabla d_i(p_j),$$

which are the $g_\pi(p_j)$ for the s reduced spatial median problems with vertices $A_j \triangleq \{a_i \mid j \in J_i\}$ and weights $\lambda_{j,i}$. Likewise $h(\bar{z},\bar{v};\bar{p}) = \sum_{j=1}^{s} h(z_j, v_j; p_j)$, where

$$h(z_j, v_j; p_j) = (\sum_{i\notin\pi(p_j)} \nabla d_i(p_j) - v_j)^T z_j + \sum_{i\in\pi(p_j)} \|z_j\|_i = g_j^T z_j + \sum_{\substack{i\in\pi(p_j) \\ J_i\ni j}} \lambda_{j,i} \|z_j\|_i \,, \tag{24}$$

which are h for the same reduced problems. It follows that $\tilde{z}$ required by Lemma 1 can be chosen independently for each j, together with v_j. (Note that v_j only depends on the $\mathcal{R}(\rho_{\pi(p_j)})$ part of z_j, i.e. $\tilde{z}_j$.) However, $S_\pi(\bar{p})$ does not split into clusters quite so well: it remains dependent on the whole original data set, $S_\pi(\bar{p}) = (S_{\pi,\mathrm{full}}(p_1),\dots,S_{\pi,\mathrm{full}}(p_j))$, where $S_{\pi,\mathrm{full}}(p_j) \triangleq \sum_{i\in\{1,\dots,n\}\setminus\pi(p_j)} S_i(p_s)$. Despite this, the direction of (6),

$$z(\bar{p},\bar{v}) = (\dots, -\bar{\rho}_{\pi(p_j)} S_{\pi,\mathrm{full}}^{\dagger}(p_j) g_j + \alpha\rho_{\pi(p_j)}\tilde{z}_j, \dots),$$

can be calculated almost independently for each j, and we in fact have:

Theorem 4. *For the multisource Weber problem, Algorithm 1 reduces to calculating at each step, for the spatial median problems*

$$\min_{p'_j} \sum_{i:j\in J_i} \lambda_{j,i} d_i(p'_j) \quad (j = 1,\dots,s), \tag{25}$$

one iteration starting from p_j, of the convex SOR-Weiszfeld algorithm, modified to use $S_{\pi,\text{full}}(p_j)$ (instead of $S_\pi(p_j)$ for the data set A_j), in the direction (6)*. If the sum is empty, the point remains unaltered. The step sizes ω and α must be the same for all j, and valid for the full problem.*

Since $S_{\pi,\text{full}} \geq S_\pi \geq 0$ (component-wise), the effect of this modification in both (6) and (7) (for the problem (25)), is to shorten the step. In our study on how the choice of step lengths affects the boundedness of the iterates, we may therefore consider the application of the unperturbed Weiszfeld algorithm (for incomplete data) without the $S_{\pi,\text{full}}$-modification, to the problems (25).

4.2 Boundedness and convergence

If we are working with complete data and step size $\omega = 1$, it is well known that each iterate of the (convex unperturbed) Weiszfeld algorithm is in the convex hull of the data points when the current iterate is not one of the vertices; cf. [16]. In fact, when an iterate equals one of the vertices, we can freely choose the step size as small as we want – the condition $\omega \geq 1$ does not apply to such points – and therefore keep things bounded. Since the convex hull of a subset of points belongs in the convex hull of the full set, we can therefore keep the iterates bounded in this case.

Similarly in our case of incomplete data, for $\pi(p) = \emptyset$ and convex problems of spatial medians, each coordinate of $T_1(p)$ is in the convex hull of the corresponding (non-missing) coordinates of the data. (We drop v from the parameters of T_ω for the convex sub-problems, since it is zero.) But our convergence theorem does not guarantee convergence for a fixed step size for all kinds of incomplete data sets. It is therefore imperative to study how the selection of step sizes affects boundedness of the iterates.

Let $\hat{p} \in \mathbb{R}^m$ be some reference point, e.g. a spatial median of the data, $L > 1$, and $\pi \triangleq \pi(p_j)$. Then, for the difference of $p'_j \triangleq T_\omega(p_j)$ and $\hat{p}$, following (10), we have for the coordinates k present in $\mathcal{R}(\bar{\rho}_\pi)$ that

$$\begin{aligned}\left|(p'_j - \hat{p})_k\right| = |((1-\omega)(p_j - \hat{p}) + \omega(c - \hat{p}))_k| &\leq |1-\omega|\,|(p_j - \hat{p})_k| + \omega\,|(c-\hat{p})_k| \\ &\leq |1-\omega|\,|(p_j - \hat{p})_k| + \omega C_k,\end{aligned}$$

with c some point in the coordinate-wise convex hull of the data (as an average of a_k weighted by S_k), and $C_k = \max_c |(c-\hat{p})_k|$. Therefore, if $|(p_j - \hat{p})_k| < (L - \omega)/(\omega - 1)C_k$ for some valid $\omega > 1$, then we have that $\left|(p'_j - \hat{p})_k\right| < LC_k$. Since for $L > 1$, $(L-\omega)/(\omega-1) \nearrow \infty$ as $\omega \searrow 1$, such an ω can always be found.

To bound $(p'_j - \hat{p})_k$ for coordinates in $\mathcal{R}(\rho_\pi)$, we alter the parameter α in the iteration. By the definition of the step $z(p)$ in (6), $\left|(p'_j - \hat{p})_k\right| \leq |(p_j - \hat{p})_k| + \alpha\omega$. By Lemma 1, the iteration is descending for each $\omega \in (1,2)$ and $\alpha \in (0, \alpha_0)$, with $\alpha_0 > 0$. We may therefore make $\alpha\omega > 0$ arbitrarily close to zero. Thus, with $|(p_j - \hat{p})_k| < LC_k - \alpha\omega$, we have $\left|(p'_j - \hat{p})_k\right| < LC_k$. Hence we can state:

Theorem 5. *With the choice of α and ω as above, the sequence of iterates for the perturbed SOR-Weiszfeld algorithm of Theorem 4 can be held bounded for the K-spatial-medians objective. In consequence, the convergence results of Theorem 2 apply. Furthermore, with choices of $\Lambda \in \mathcal{W}_{\text{ext}}$, we can take $\mathcal{D} = \mathcal{D}_N$.*

Proof. Above we have derived upper bounds for ω and α for each cluster centre to stay in the box $(p - \hat{p}) + \prod_{k=1}^m(-LC_k, LC_k)$ for arbitrary $L > 1$ and reference point $\dot{p}$, if the previous iterates satisfy this. Because the number of conditions is finite, and allow for ω to vary in some non-singleton range above and including 1, there's enough leeway for ω for it to be altered in such a manner that the conditions in Theorem 2 on ω are met. Furthermore, the K-spatial-medians objective function clearly is bounded from below, so the theorem applies.

In the choice (23) of v_j used to obtain (24), we choose $W_i^2\tilde{z}/\|\tilde{z}\|_i \in \partial d_i(p_j)$ for $i \in \pi(p_j)$. These are in the limit of gradients of differentiable points of $-f_2^{\mathrm{KM}}$, for at these points $\nabla d_i(p_j)$ takes the form $W_i^2(p_j - a_i)/d_i(p_j)$. Furthermore, directions in $\mathcal{D}_{\mathrm{N}}(-f^2_{\mathrm{KM}})(\bar{p})$ have $\Lambda \in \mathcal{W}_{\mathrm{ext}}$. For, if $\#J_i(\bar{p}) > 1$, then $-f_2^{\mathrm{KM}}$ is not differentiable, and hence at differentiable points $\mathcal{W} = \mathcal{W}_{\mathrm{ext}}$. As directions in $\mathcal{D}_{\mathrm{N}}$ are limits of directions at differentiable points, by the preceding we must have have $\Lambda \in \mathcal{W}_{\mathrm{ext}}$ for such directions. Now, if $\bar{p}$ is $\mathcal{D}_{\mathrm{N}}$-critical, then there's a choice of weights $\Lambda \in \mathcal{W}_{\mathrm{ext}}$ for which $(\bar{p}, \Lambda)$ solves (25) for each j. Therefore, with such choice of Λ, $\bar{v} \in \mathcal{D}_N(\bar{p})$. □

4.3 Optimality

Extend Λ by setting $\lambda_{j,i} = 0$ for $j \notin J_i$. For fixed Λ, we may then reformulate the objective of Theorem 4 in a combined form as finding $\bar{p}'$ such that $F(\bar{p}'; \Lambda) < F(\bar{p}; \Lambda)$ for the function

$$F(\bar{p}; \Lambda) \triangleq \sum_j \sum_i \lambda_{j,i} d_i(p_j). \tag{26}$$

Theorem 6. *The point $\bar{p}^*$ is a local minimum of* (20) *if and only if it minimises $F(\cdot; \Lambda)$ for all $\Lambda \in \mathcal{W}(\bar{p}^*)$.*

Proof. Necessity is obvious: $(f_1 + f_2^{\mathrm{KM}})(\bar{p}) = \sum_{i=1}^s \min_j d_i(p_j) \leq F(\bar{p}; \Lambda)$ with equality at $\bar{p}^*$, for all $\Lambda \in \mathcal{W}(\bar{p}^*)$. Hence if $\bar{p}^*$ isn't a minimiser of the convex function $F(\cdot; \Lambda)$ for some such Λ, it can't minimise (20) even locally.

As for sufficiency: for all $\bar{p}$ sufficiently close to $\bar{p}^*$, $\mathcal{W}(\bar{p}) \subset \mathcal{W}(\bar{p}^*)$ (with the identification $\lambda_{j,i} = 0$ for $j \notin J_i$). Therefore, sufficiently close to $\bar{p}^*$, by the definition of $\mathcal{W}(\bar{p})$, $f_1(\bar{p}) + f_2^{\mathrm{KM}}(\bar{p}) = \min\{F(\bar{p}; \Lambda) \mid \Lambda \in \mathcal{W}(\bar{p})\} \geq \min\{F(\bar{p}; \Lambda) \mid \Lambda \in \mathcal{W}(\bar{p}^*)\}$. But since $F(\cdot; \Lambda)$ is minimised at $\bar{p}^*$ for all $\Lambda \in \mathcal{W}(\bar{p}^*)$, it must be a local minimiser of $f_1 + f_2^{\mathrm{KM}}$ as well. □

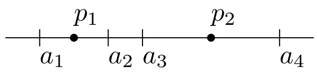
Corollary 1. *i) If $\#J_i(\bar{p}^*) = 1$ for all $i = 1, \ldots, n$, and $\bar{p}^*$ minimises $F(\cdot; \Lambda^*)$ for the unique $\Lambda^* \in \mathcal{W}(\bar{p}^*)$, then $\bar{p}^*$ is a local minimiser of* (20)*. ii) If $\#J_i(\bar{p}) > 1$, and we have $\pi(p_j) = \emptyset$ for some $j \in J_i(\bar{p})$, then $\bar{p}$ is not a local minimiser.*

Proof. The first claim is obvious from the preceding theorem. As for the second claim, suppose $\bar{p}$ minimises $F(\cdot; \Lambda^*)$ for some $\Lambda^* \in \mathcal{W}(\bar{p})$. Let $j, j' \in J_i(\bar{p})$, $j \neq j'$, and $\pi(p_j) = \emptyset$. Let Λ be altered from Λ^* by moving weight between $\lambda_{j,i}$ and $\lambda_{j',i}$. This will not change the value of F at $\bar{p}$. However, the condition $0 \in \{\nabla \sum_i \lambda_{j,i} d(a_i, p_j)\}$ will be upset, and hence the value of $f_1 + f_2^{\mathrm{KM}}$ can be improved locally. □

Corollary 2. *If $\bar{p}^*$ is $\mathcal{D}_{\mathrm{N}}$-critical and $\#J_i(\bar{p}^*) = 1$ for all i, then $\bar{p}^*$ is a local solution of* (20).

Proof. The condition $\#J_i(\bar{p}^*) = 1$ forces Λ^* to be unique. Therefore, also $(1 - \lambda_{j,i})\partial d_i(p_j^*)$ reduces to the singleton $\{0\}$ in (22). Hence $\bar{v}(\bar{p}^*)$ is uniquely determined. It then follows from $\mathcal{D}_N$-criticality that $\bar{p}^*$ minimises (25) for all j, and consequently minimises (26). That $\bar{p}^*$ is a local solution follows from Corollary 1. □

Remark 4. In fact, that $\#J_i(p^*) = 1$ or $\min_j d_i(p_j^*) > 0$ for all i forces $\bar{v}(\bar{p}^*)$ to be uniquely determined by Λ. We may show that such points are in fact critical and not only semi-critical. However, a simple example on the real line furnishes that the relaxed condition does not guarantee local optimality:



Here p_1 and p_2 are at equal distance from a_3. If a_3 is assigned to the cluster of p_2, we have a critical point, yet assignment to p_1 shows that both cluster centres can be improved by just a small move of either or both p_1 or p_2 to the right.

Corollary 3. *Under conditions of Theorem 5, with choices of $\Lambda \in \mathcal{W}_{\text{ext}}(\bar{p})$, if the iterates $\{\bar{p}_r\}$ of the algorithm of Theorem 4 converge to $\bar{p}^*$, then it is either a local minimiser, or has disputed vertices: $\#J_i(\bar{p}^*) > 1$ for some $i \in \{1, \dots, n\}$.*

Proof. Since $\{\bar{p}_r\}$ converge to $\bar{p}^*$, if $\{\bar{v}(\bar{p}_r)\}$ diverges, then $-f_2^{\text{KM}}$ is non-differentiable at $\bar{p}^*$ (cf. Remark 1(i)). This says that there are disputed vertices. If $\{\bar{v}(\bar{p}_r)\}$ also converges, then by Theorem 5 (and Theorem 2), $\bar{p}^*$ is $\mathcal{D}_{\text{N}}$-critical, and the previous corollary applies. □

Remark 5. Suppose that eventually in the method, the assignments of vertices to clusters is unique. Then, if the data set is complete (or more generally $\#\pi(p_i) \leq 1$ always), we have convergence to the set of local minimisers (being able to analyse the method on each cluster separately, applying Remark 1(iii)). Therefore, with such simple data, non-convergence is always a case of dispute over assignment of vertices to clusters.

4.4 Discussion and multiobjective interpretation

We have thus provided a method for the multisource Weber problem, providing convergent sequences to semi-critical points of the problem and often, in fact, to local minima. Our method does not depend on solving s inner spatial median problems (likely with the Weiszfeld algorithm) between *each* step of allocating vertices to clusters. Instead, we only solve a single perturbed spatial median problem, which amounts to running s "tilted" SOR-Weiszfeld iterations in parallel, with tilts calculated from the results of all the s previous iterations, as was explained in Section 3.2.

If we choose $\{\lambda_{j,i}\}$ as extreme points of the feasible sets, then in some sense, our method is "dual" to the basic K-means type algorithm: in that algorithm, spatial medians are calculated between assignments of vertices to clusters, whereas in our method vertices are assigned to clusters between iterations of a method to find spatial medians. To summarise, Algorithm 1 reduces to the following:

Algorithm 2 (K-means type method with single step SOR-Weiszfeld)**.**

1. Choose some starting points p_j $(j = 1, \dots, s)$.
2. Assign each vertex a_i $(i = 1, \dots, n)$ to one of the clusters A_j corresponding to closest p_j $(j = 1, \dots, s)$.
3. To obtain p_j', calculate for the (convex) spatial median problem on A_j, one iteration of Algorithm 1 with the modified direction

$$z^{\text{KM}}(p_j) \triangleq -\bar{\rho}_{\pi(p_j)} S^{\dagger}_{\pi,\text{full}}(p_j) g_{\pi}(p_j) + \alpha \rho_{\pi(p_j)} \tilde{z}_j, \tag{27}$$

   where $g_{\pi}(p_j)$ and $\tilde{z}_j$ are calculated for the data A_j. See below for constraints on step sizes.
4. Continue from step 2 unless a stopping criterion is satisfied.

The step lengths $\omega \in [1, 2)$ and α should be the same for each cluster, according to Theorem 4. Since (7) defining the bound α_2 for the whole problem, is the sum of $S_{\pi,\text{full}}$-modified conditions for the sub-problems, it suffices to to bound α from above by the minimum of the upper bounds for the sub-problems. Theorem 2 sets some minor restrictions on $\omega \in [1, 2)$ to avoid oscillation. Theorem 5 sets additional upper bounds on the step lengths by the coordinate-wise bound $LC_k > C_k$ on $\left|(\dot{p}'_j)_k\right|$, which we may, however, choose arbitrarily large.

Example 2. When $W_k = w_k I$ for all $k = 1, \ldots, n$, $S_{\pi,\text{full}}$ is proportional to the identity; cf. Example 1. Therefore, in that case, (27) is simply a shortened standard Weiszfeld step for the data A_j. The effect of the data outside the cluster A_j is therefore to damp too quick convergence to its centre. For more complex weights W_k, the same conclusion holds coordinate-wise.

In light of the multiobjective clustering criteria considered in Section 3, it is interesting to interpret the K-spatial-medians as one scalarisation of a more general problem

$$\min_{\bar{p}} (f_1, f_2^{\text{KM}})(\bar{p}).$$

The meaning of the objective f_1 is the same as before. What the objective f_2^{KM} means is: place all but the closest cluster centre as far from a_i as possible. This sounds like a very natural criteria. Thus, it will be interesting to look at the results of $\min_{\bar{p}} f_1(\bar{p}) + \lambda f_2^{\text{KM}}(\bar{p})$ for $\lambda \in [0, 1]$.

For $\lambda \in [0, 1)$, Theorem 7 is applicable to proving boundedness of the level sets. To see this, consider the inclusions $\lambda \mathcal{R}(\partial(-f_2^{\text{KM}})) \subset \lambda \bigcup_\Lambda \mathcal{R}(\partial f_\Lambda) \subset \operatorname{int} \mathcal{R}(\partial f_1)$, where $f_\Lambda : \bar{p} \mapsto \sum_{j=1}^s \sum_{i=1}^n (1 - \lambda_{j,i}) d_i(p_j)$, and Λ ranges over all the admissible weights $\{\lambda_{j,i}\}$ with $\lambda_{j,i} \geq 0$ and $\sum_i \lambda_{j,i} = 1\}$. The first inclusion can be seen from taking the union over $\bar{p}$ in the expression (21). To see the second inclusion, note that $f_1 = f_{\Lambda_1}$ for Λ_1 with all zero weights. Therefore $f_1 - \lambda f_\Lambda$ is a convex function with bounded level sets for $\lambda \in [0, 1)$ and admissible Λ. Thus an application of Theorem 7 yields that $\lambda \mathcal{R}(\partial f_\Lambda) \subset \operatorname{int} \mathcal{R}(\partial f_1)$. Finally since the inclusions above hold for some other $\lambda' \in (\lambda, 1)$, the result must hold for the closure as well, i.e. $\operatorname{cl} \mathcal{R}(\partial(-\lambda f_2^{\text{KM}})) \subset \operatorname{int} \mathcal{R}(\partial f_1)$. Now Theorem 7 applies again.

5 Experiments

In this section we present some experiments with the proposed algorithm(s) and clustering formulations. It is not our intent to provide thorough statistically significant testing and comparison of the method with alternatives, but rather to provide minimal experimental proof that the method works, and to visually compare the KM and MO clustering objectives. Especially, the statistical and computational properties of the K-spatial medians, along with significant amount of tests with real and simulated data, are covered in [32]. Our forthcoming paper [30] will also include more extensive computational results with Algorithm 1 in other applications, related to the MO clustering objective.

Figures 1 and 2 show results for two cases using both the problem of Section 3.2 (MO), and the multi-objective formulation of the K-spatial-medians (KM) discussed in Section 4. The number of clusters is three, and the total number of vertices is 90. The weight λ was randomly varied between zero and the indicated upper limit for 30 samples in each case. The stopping criterion used was $\max_{j=1,2,3} \left\|p'_j - p_j\right\| < 10^{-6}$ and the maximum number of iterations was 300. The

nits: 57/42/27/191
(a) MO, $\lambda_{\text{sup}} = 90/4$

nits: 29.6/30.5/15/78
(b) KM, $\lambda_{\max} = 1$

Figure 1: Results for a task with three clear clusters with varying λ and $\omega = 1.5$.

actual mean, median, minimum and maximum numbers of iterations (nits) of the perturbed SOR-Weiszfeld method to reach the threshold is given in the figures (in that order: mean/median/min/max). The bigger dots in the figures denote the data, and the smaller dots the clusters centres.

As we can see, for $\lambda = 0$ the result is in both cases the spatial median of the data. From there, the solutions continuously move towards the centres of clusters, as λ varies towards the respective upper bound for λ ($\lambda_{\text{sup}} = (n/2)/(s-1)$ for MO, from Theorem 3, and $\lambda_{\max} = 1$ for KM, from the analysis of Section 4.2), just as suggested by results of sensitivity analysis of optimisation problems under some assumptions on the second-order behaviour of the objective function at the solution; cf. e.g. [3]. Interestingly, the paths the solutions travel are very similar for both KM and MO, and the paths for MO pass closely to the cluster centres for KM, but "overshoot" slightly for big λ. This resemblance isn't entirely unexpected, however: for tightly packed clusters, we should have $d(p_k^*, p_j^*) \approx \sum_{i:k\in J_i} d(a_i, p_j^*)/\#\{i : k \in J_i\}$ for all $k \neq j$. In case of the K-spatial-medians, the small amount of total iterations used is also noteworthy when compared to the basic K-means-type algorithm, where a comparable number of iterations would be used in the inner (SOR-Weiszfeld) algorithm used to calculate the spatial medians [18]. One may also note that the MO formulation has required more iterations in our tests. But since this number is dependent on the stopping criterion, and absolute quality of the solutions is not known, not much conclusions can be drawn.

nits: 68.8/63/39/151
(a) MO, $\lambda_{\text{sup}} = 90/4$

nits: 43.4/39/18/132
(b) KM, $\lambda_{\text{max}} = 1$

Figure 2: Results for a task with three less clear clusters with varying λ and $\omega = 1.5$.

6 Miscellaneous applications and final remarks

The Weber problem with attraction and repulsion studied in [6] is also a problem of perturbed spatial median. In this problem, some of the weights w_i are allowed to be negative, creating repulsive points and making the problem diff-convex. The problem is also studied on the plane in [11], where another modification of the Weiszfeld algorithm is developed.

It is also interesting to note the superficial similarity of the scalarisation of our bi-objective clustering formulation to the Euclidean multifacility location problem. In the latter problem, the f_2 component has sign changed from ours, aside from including weights for all the distances, needed for the problem to not reduce to a single-facility problem. Our algorithm is therefore not directly suited for solving this problem, yet a Weiszfeld-type algorithm can be derived for it; see [21].

As for the applicability of our algorithms, we did manage to characterise the local minimisers of the multisource Weber problem in such a manner that our algorithm can be seen to converge to local minimisers in a wide variety of cases. For our multiobjective clustering formulation such an easily "interpretable" characterisation of local minima has been elusive, however. Nevertheless, our limited numerical experiments suggest potentially good clustering behaviour.

Acknowledgements The authors would like to thank professors Marc Teboulle, Dan Tiba, and Marko Mäkela for feedback on an earlier version of the manuscript.

References

[1] S. Arora, P. Raghavan, and S. Rao, *Approximation schemes for Euclidean*

k-medians and related problems, in ACM Symposium on Theory of Computing, 1998, pp. 106–113.

[2] I. Bongartz, P. H. Calamai, and A. R. Conn, *A projection method for l_p norm location-allocation problems*, Mathematical Programming, 66 (1994), pp. 283–312.

[3] J. F. Bonnans and A. Shapiro, *Optimization problems with perturbations: A guided tour*, SIAM Review, 40 (1998), pp. 228–264.

[4] J. Brimberg, P. Hansen, N. Mladenović, and E. D. Taillard, *Improvements and comparison of heuristics for solving the uncapacitated multisource Weber problem*, Operations Research, 48 (2000), pp. 444–460.

[5] J. Brimberg and N. Mladenović, *Degeneracy in the multi-source Weber problem*, Mathematical Programming, 85 (1999), pp. 213–220.

[6] P.-C. Chen, P. Hansen, B. Jaumard, and H. Tuy, *Weber's problem with attraction and repulsion*, Journal of Regional Science, 32 (1992), pp. 467–486.

[7] ——, *Solution of the multisource Weber and conditional Weber problems by D.-C. programming*, Operations Research, 46 (1998), pp. 548–562.

[8] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Canadian Mathematical Society series in mathematics, Wiley-Interscience, 1983.

[9] L. Cooper, *Heuristic methods for location-allocation problems*, SIAM Review, 6 (1964), pp. 37–53.

[10] D. R. Cox, *Note on grouping*, Journal of the American Statistical Association, 52 (1957), pp. 543–547.

[11] Z. Drezner and G. O. Wesolowsky, *The Weber problem on the plane with some negative weights*, INFOR, 29 (1991), pp. 87–99.

[12] M. Dür, *A parametric characterization of local optimality*, Mathematical Methods of Operations Research, 57 (2003), pp. 101–109.

[13] J.-B. Hiriart-Urruty, *Conditions for global optimality*, in Handbook of Global Optimization, R. Horst and P. M. Pardolos, eds., Kluwer Academic Publishers, 1995, pp. 1–26.

[14] ——, *Conditions for global optimality 2*, Journal of Global Optimization, 13 (1998), pp. 349–367.

[15] H. Horst and N. V. Thoai, *DC programming: Overview*, Journal of Optimization Theory and Applications, 103 (1999), pp. 1–43.

[16] H. W. Kuhn, *A note on Fermat's problem*, Mathematical Programming, 4 (1973), pp. 98–107.

[17] T. Kärkkäinen and S. Äyrämö, *Robust clustering methods for incomplete and erroneous data*, in Proceedings of the Fifth Conference on Data Mining, WIT Press, 2004, pp. 101–112.

[18] ——, *On computation of spatial median for robust data mining*, in Proceedings of Sixth Conference on Evolutionary and Deterministic Methods for Design, Optimisation and Control with Applications to Industrial and Societal Problems (EUROGEN 2005), R. Schilling, W. Haase, J. Periaux, and H. Baier, eds., Munich, Germany, September 2005, FLM, TU Munich.

[19] K. Miettinen, *Nonlinear Multiobjective Optimization*, Kluwer Academic Publishers, Boston, 1999.

[20] M. Mäkelä and P. Neittaanmäki, *Nonsmooth Optimization*, World Scientific Publishing, Singapore, 1992.

[21] F. Radó, *The Euclidean multifacility location problem*, Operations Research, 36 (1988), pp. 485–492.

[22] R. T. Rockafellar, *Level sets and continuity of conjugate convex functions*, Transactions of the American Mathematical Society, 123 (1966), pp. 46–63.
[23] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1972.
[24] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, 1998.
[25] S. Z. Selim and M. A. Ismail, *K-means-type algorithms: A generalized convergence theorem and characterization of local optimality*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6 (1984).
[26] M. Teboulle, *A unified continuous optimization framework for center-based clustering methods*, Journal of Machine Learning Research, 8 (2007), pp. 65–102.
[27] H. Tuy, *D.C. optimization: Theory, methods and algorithms*, in Handbook of Global Optimization, R. Horst and P. M. Pardolos, eds., Kluwer Academic Publishers, 1995, pp. 149–216.
[28] T. Valkonen, *Convergence of a SOR-Weiszfeld type algorithm for incomplete data sets*, Numerical Functional Analysis and Optimization, 27 (2006), pp. 931–952. (Errata in vol. 29, no 9–10, 2008).
[29] ——, *An errata to: Convergence of a SOR-Weiszfeld type algorithm for incomplete data sets*, Numerical Functional Analysis and Optimization, 29 (2008), pp. 1201–1203.
[30] T. Valkonen and T. Kärkkäinen, *Continuous reformulations and heuristics for the Euclidean travelling salesperson problem*, ESAIM: Control, Optimization and Calculus of Variations, (2008). published online (E-first).
[31] E. Weiszfeld, *Sur le point pour lequel la somme des distances de n points donnés est minimum*, Tôhoku Mathematics Journal, 43 (1937), pp. 355–386.
[32] S. Äyrämö, *Knowledge Mining Using Robust Clustering*, no. 63 in Jyväskylä Studies in Computing, University of Jyväskylä, 2006. Ph.D Thesis.

A Boundedness of level sets of DC functions

In this section some conditions are provided for the boundedness of level sets of differences of convex functions (DC functions). These are functions of the form $f_\nu \triangleq f - \nu$ with both f and ν convex. See, for example, the survey papers [27, 15] for an overview of results related to optimisation theory for such functions, and application examples. For some optimality conditions – that unfortunately do not seem to provide a simple characterisation of optimality for our application examples – see [13, 14, 12].

We will denote the range of the subgradient of a convex function f by $\mathcal{R}(\partial f) \triangleq \bigcup_{p\in\mathbb{R}^m} \partial f(p)$, which is non-empty for proper convex functions. The convex conjugate is denoted by $f^*(\xi) \triangleq \sup_{p\in\mathbb{R}^m}\{\xi^T p - f(p)\}$.

Lemma 10. *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a closed proper convex function. Suppose that for some $p \in \operatorname{dom} f$, for some $v \in \partial f(p)$, we have $v \in \operatorname{bd}\mathcal{R}(\partial f)$. Then $\partial f(p + \alpha z) \subset \operatorname{bd}\mathcal{R}(\partial f) \cap \mathcal{R}(\partial f) \cap (v + \{z\}^\perp)$ for all $\alpha > 0$ and $z \in N_{\operatorname{cl}\mathcal{R}(\partial f)}(v) \setminus \{0\}$.*

Proof. The set $\operatorname{cl}\mathcal{R}(\partial f)$ is convex (see e.g. [23, Section 24]). Set $p' = p + \alpha z$ for $z \in N_{\operatorname{cl}\mathcal{R}(\partial f)}(v) \setminus \{0\}$, and suppose $v' \in \partial f(p')$. (If $\partial f(p') = \emptyset$, the claim of the lemma is immediate.) Then by the monotonicity of ∂f as the subgradient of a convex function, we have $(p' - p)^T(v' - v) \geq 0$. That is, $z^T(v' - v) \geq 0$. This says that either $v' \notin \operatorname{cl}\mathcal{R}(\partial f)$ or $z^T(v' - v) = 0$. Since the former is by definition not possible, v' must lie in the plane orthogonal to z anchored at v. Because z is normal to the convex set $\operatorname{cl}\mathcal{R}(\partial f)$ at v, it must be that $v' \notin \operatorname{int}\mathcal{R}(\partial f)$. Because $v' \in \partial f(p') \subset \mathcal{R}(\partial f)$, the claim of the lemma follows. □

From the assumption in the following theorem, that $\mathcal{R}(\partial f)$ bounded, it follows of course that f is finite-valued, so that the difference $f_\nu = f - \nu$ is also pointwise well-defined, even if ν were not finite-valued.

Theorem 7. *Suppose that f and ν are closed proper convex functions in $\mathbb{R}^m$, with $\mathcal{R}(\partial f)$ bounded. For the level sets $\operatorname{lev}_c f_\nu \triangleq \{p \mid f_\nu(p) \le c\}$ to be bounded, it is sufficient that $\operatorname{cl}\mathcal{R}(\partial\nu) \subset \operatorname{int}\mathcal{R}(\partial f)$ and necessary that $\mathcal{R}(\partial\nu) \subset \operatorname{int}\mathcal{R}(\partial f)$.*

Proof. Let $A \triangleq \mathcal{R}(\partial f)$ and $C \triangleq \mathcal{R}(\partial\nu)$.

First we tackle sufficiency. We may assume that $0 \in \operatorname{int} A$, because if the interior is empty, the required condition cannot hold, and for arbitrary $\xi \in \operatorname{int} A$, we may rewrite $(f - \nu)(p) = (f(p) - \xi^T p) - (\nu(p) - \xi^T p)$, yielding another DC representation of the same function f_ν, for which $0 \in \mathcal{R}(\partial(f - \xi^T))$, and the required inclusion condition holds. Likewise we may assume that $\nu(0)$ is finite.

Recall that $\tilde{\nu}_q^v(p) \triangleq \nu(q) + v^T(p-q)$. Since $\nu(p) = \sup_{q\in\mathbb{R}^m, v\in\partial\nu(q)} \tilde{\nu}_q^v(p)$, with the supremum achieved (at least by $q = p$), we may expand

$$\begin{aligned}\operatorname{lev}_c f_\nu &= \{p \mid f(p) - \sup_{q,v} \tilde{\nu}_q^v(p) \le c\} \\ &= \{p \mid \inf_{q,v}(f(p) - \tilde{\nu}_q^v(p)) \le c\} \\ &= \bigcup_{q\in\mathbb{R}^m, v\in\partial\nu(q)} \{p \mid f(p) - \nu(q) - v^T(p-q) \le c\}.\end{aligned}$$

But, since $v \in \partial\nu(q)$, we have $\nu(0) - \nu(q) \ge v^T(0 - q)$, or that $\nu(q) - v^T q \le \nu(0)$. Hence q can be removed from the equation, and we have

$$\operatorname{lev}_c f_\nu \subset \bigcup_{v\in C} \{p \mid f(p) - v^T p \le c_0\} = \bigcup_{v\in C} \operatorname{lev}_{c_0}(f - v)$$

with $c_0 = c + \nu(0)$. Therefore it suffices to prove that the sets $\operatorname{lev}_c(f - v)$ are uniformly bounded over v for any fixed c. Boundedness of $\operatorname{lev}_c(f - v)$ when $v \in \operatorname{int} A$ is well known; see e.g. [22]. For the uniform boundedness of this family of sets a little more work is needed.

By the inclusion $\operatorname{cl} C \subset \operatorname{int} A$, $0 \in A$ and $\operatorname{cl} A$ being convex (cf. [23, Section 24]) and bounded, every $v \in \operatorname{cl} C \setminus \{0\}$ has an $\epsilon_v \in (0, 1/4)$ such that $B(v, 4\epsilon_v) \subset \operatorname{int} A$ and $v/(1 - 4\epsilon_v) \in \operatorname{cl} A$. Since $\operatorname{cl} C$ is a subset of the bounded set A, it is compact, and we can find a finite set $C^* \subset \operatorname{cl} C \setminus \{0\}$ such that the sets $v^* + 2\epsilon_{v^*} A$ for $v^* \in C^*$ cover C. It then suffices to prove that each of the sets $L_{v^*} \triangleq \bigcup_{v\in v^* + 2\epsilon_{v^*} A} \operatorname{lev}_c(f - v)$ is bounded for $v^* \in C^*$, which are finite in number.

To prove this, first notice that for any $p \in \mathbb{R}^m$,

$$\left|(f(p) - (v^*)^T p) - (f(p) - v^T p)\right| = \left|(v^* - v)^T p\right|.$$

But

$$\sup_{v\in v^* + 2\epsilon_{v^*} A} \left|(v^* - v)^T p\right| = \sup_{z\in A} 2\epsilon_{v^*} \left|z^T p\right| = 2\epsilon_{v^*} \left|z^*(p)^T p\right|$$

for some $z^*(p)$ on the border of A. Therefore, for $v \in v^* + 2\epsilon_{v^*} A$,

$$L_{v^*} \subset \{p \mid f(p) - (v^*)^T p \le c + 2\epsilon_{v^*} \left|z^*(p)^T p\right|\}.$$

But as $v^*/(1-4\epsilon_{v^*}) \in \operatorname{cl} A$ by our choice of ϵ_{v^*}, it holds that

$$\left|(v^*)^T p\right| = (1-4\epsilon_{v^*}) \left|(v^*/(1-4\epsilon_{v^*}))^T p\right| \leq (1-4\epsilon_{v^*}) \left|z^*(p)^T p\right|,$$

and

$$L_{v^*} \subset \{p \mid f(p) \leq c + (1-2\epsilon_{v^*}) \left|z^*(p)^T p\right|\}.$$

We must still bound f from below. For this, notice that

$$\begin{aligned} f(p) &= \sup\{z^T p - f^*(z) \mid z \in \mathbb{R}^m\} \\ &\geq \sup\{z^T p - f^*(z) \mid z \in A'\} \\ &\geq (1-\epsilon_{v^*}) \left|z^*(p)^T p\right| - \sup\{f^*(z) \mid z \in A'\} \end{aligned}$$

for $A' = (1-\epsilon_{v^*})(\operatorname{int} A) \subset \operatorname{int} A$. Thus, if f^* is bounded within A' by c', we get

$$L_{v^*} \subset \{p \mid \epsilon_{v^*} \left|z^*(p)^T p\right| \leq c + c'\},$$

and this is clearly bounded, because we have assumed $0 \in \operatorname{int} A$, whence $\left|z^*(p)^T p\right| \geq \delta \|p\|$ for some $\delta > 0$.

To prove the boundedness of f^* within A', we note that the interior of the finite domain of f^* is contained in $\operatorname{int} A$; see [23, Section 24] again. Hence if f^* wasn't bounded in A', a bounded set, we could find a sequence $\{v_i\}_{i=1}^{\infty} \subset A'$ converging to some $v \in \operatorname{bd} A'$ for which $f^*(v) = \infty$. But this contradicts the finiteness of f^* on $\operatorname{int} A$.

As for necessity of $C \subset \operatorname{int} A$, suppose first that for some for some q and $v \in \partial\nu(q)$, $v \notin A$. Then $v \notin \partial f(p)$ for any $p \in \mathbb{R}^m$, i.e. $0 \in \partial(f - v)$ has no solution. Therefore $f - v$ must be descending in some direction p for infinitely large values of $\|p\|$. Since $f - \nu \leq f - \tilde{\nu}_q^v$, it follows that $f - \nu$ must have unbounded level sets.

Let then $v \in \operatorname{bd} A \cap A$, and $v \in \partial f(q)$ for some q. By Lemma 10, since the normal cone at the border of a convex set contains non-zero elements, it follows that there exists a direction z such that for $q_r \triangleq q + rz$, there exists $v_r \in \operatorname{bd} A \cap \partial f(q_r)$ such that $(v_r - v)^T z = 0$. But then

$$\begin{aligned} f(q_r) - v^T q_r &\leq f(q) - v_r(q - q_r) - v^T q_r \\ &= f(q) + r v_r^T z - v^T q - r v^T z \\ &= f(q) - v^T q. \end{aligned}$$

The values of $f - v$ along the line $\alpha \mapsto q + \alpha z$ are thus bounded by $f(q) - v^T q$ and therefore $f - v$ and consequently f_ν must have unbounded level sets. □

To see that $\operatorname{cl} \mathcal{R}(\partial\nu) \subset \operatorname{int} \mathcal{R}(\partial f)$ is not necessary, consider the real functions $f : p \mapsto |p|$ and

$$\nu : p \mapsto \sup_{k=1,2,3,\ldots} \nu_k(p) \quad \text{with} \quad \nu_k(p) = \sum_{i=1}^{k} 2^{-i}(|p| - 2^i). \tag{28}$$

Then $\mathcal{R}(\partial f) = [-1, 1]$ and $\mathcal{R}(\partial\nu) = (-1, 1)$. But,

$$f(p) - \nu_k(p) = \sum_{i=1}^{\infty} 2^{-i} |p| - \nu_k(p) = \sum_{i=k+1}^{\infty} 2^{-i} |p| + \sum_{i=1}^{k} 1$$

and $f(p) - \nu(p) = \min_k(f(p) - \nu_k(p)) = f(p) - \nu_\ell(p)$ with $\ell = \max\{k \mid 2^k \leq |p|\}$, as $(f(p) - \nu_k(p)) - (f(p) - \nu_{k+1}(p)) = 2^{-k-1}|p| - 1 \leq 0$, when $|p| \leq 2^{k+1}$. Therefore, $f(p) - \nu(p) > \sum_{i=1}^k 1 = k$ for sufficiently large $|p|$. Thus, the level sets are bounded.

To see that $\mathcal{R}(\partial\nu) \subset \operatorname{int}\mathcal{R}(\partial f)$ is not sufficient, one only needs to consider f with open $\mathcal{R}(\partial f)$, and set $\nu = f$. One example of such a function is the ν of (28).

Regarding conditions on f, we have the following extension:

Corollary 4. *For proper convex functions f and ν such that the difference $f - \nu$ is well-defined, the level sets $\operatorname{lev}_c f_\nu$ are bounded if ν is closed, $\mathcal{R}(\partial\nu)$ bounded, and $\operatorname{cl}\mathcal{R}(\partial\nu) \subset \operatorname{int}\mathcal{R}(\partial f)$.*

Proof. Let $A \subset \operatorname{int}\mathcal{R}(\partial f)$ be a bounded set such that $\operatorname{cl}\mathcal{R}(\partial\nu) \subset A$, and approximate f from below by $\tilde{f}(p) \triangleq \sup\{f(q) + \xi^T(p-q) \mid \xi \in A \cap \partial f(q)\}$. $\mathcal{R}(\partial\tilde{f})$ is then bounded, and the previous theorem yields that $\tilde{f} - \nu$ and therefore also $f - \nu \geq \tilde{f} - \nu$ has bounded level sets. □

Similar conclusions do not necessarily follow if $\mathcal{R}(\partial\nu)$ is unbounded. This can be illustrated by considering the functions $p \mapsto \alpha p^2$ for varying $\alpha \in \mathbb{R}$. The difference of functions in this class is still a function in this class, and for $\alpha \leq 0$ the level sets are unbounded.

The next lemma is also of use for the verification of the assumptions of Theorem 2.

Lemma 11. *Let f and ν be proper convex functions, such that $f - \nu$ is well-defined. If $f - \nu$ has some bounded level set, it is bounded from below.*

Proof. Let A be that bounded level set. We may assume that it is non-empty, for otherwise there's nothing to prove. Then f is bounded from below on A, for otherwise it could not be proper. But ν must also be bounded from above on A, for otherwise it would attain the value $+\infty$ on some half-line starting from the border of A. Then $f - \nu$ would also have to attain $-\infty$ on this line to be well-defined, which would contradict the boundedness of A. Therefore, $f - \nu$ is bounded from below on A and consequently on all of $\mathbb{R}^m$. □

B Calculating $\hat{z}$ for non-partially-overlapping ρ_k

As noted in Section 2, we are concerned with finding the $\hat{z} \in Z(p)$ (we omit the point p from notation in this section) that minimises $h(z, v; p)$, that is, solves

$$\min_{z \in Z(p)} \left(g^T z + \sum_{k \in \pi} \|z\|_k \right) \tag{29}$$

for arbitrary $g \in \mathbb{R}^m$ in a special case. This is the case when $W_k = w_k \rho_k$ for some $w_k > 0$ and a zero-one diagonal matrix w_k, and such that the ρ_k do not "overlap" only partially. To define this notion, we introduce the notation $A \sqsubset B$ for $B - A$ being positive definite. Equivalently, in case of the ρ-matrices, $\sqsubset$ is set inclusion of the coordinates on with 1-entries on the diagonal. We also denote by $\rho \sqsubset_! \rho'$ the strict ordering $\rho \sqsubset \rho', \rho \neq \rho'$.

Now, there are said to be no partially overlapping ρ_k, if for all $k, i \in \pi$, one of the following holds: $\rho_k \rho_i = 0$, $\rho_k \sqsubset \rho_i$, or $\rho_i \sqsubset \rho_k$. These constraints are satisfied in cases like $\rho_k = \operatorname{diag}(1,1,0)$, $\rho_i = \operatorname{diag}(0,1,0)$, as well as $\rho_k = \operatorname{diag}(1,0,0)$, $\rho_k = \operatorname{diag}(0,0,1)$, but are not satisfied in cases like $\rho_k = \operatorname{diag}(1,1,0)$, $\rho_i = \operatorname{diag}(0,1,1)$.

To start solving (29), we need to do some partitioning of the coordinate ranges. Thus, let ψ be the set of maximal elements of the set of operators

$$\{\rho \mid \rho\rho_k = \rho \text{ or } \rho\rho_k = 0 \text{ for all } k \in \pi, \rho\rho_\pi = \rho\}.$$

Then

$$\hat{z} = -\sum_{\rho\in\psi} \beta_\rho g_\rho \tag{30}$$

for some $\beta_\rho \geq 0$ and $g_\rho \triangleq \rho g$; see [28] for a more detailed argument.

We denote by $\hat{\rho}_k$ the orthogonal projection into $\mathcal{R}(\rho_k) \setminus \bigcup_{\rho_i \sqsubset_! \rho_k} \mathcal{R}(\rho_i)$, and abbreviate $\hat{\beta}_k \triangleq \beta_{\hat{\rho}_k}$. Then $\psi = \{\hat{\rho}_k \mid k \in \pi\}$, and $\hat{\rho}_k$ corresponds to the fields present in ρ_k, but not in any $\rho_i \sqsubset_! \rho_k$.

Lemma 12. *Suppose ρ_τ is maximal (in $\sqsubset$). If $\hat{\beta}_\tau > 0$ and $\|\rho_\tau \hat{z}\| > 0$, then $\hat{\beta}_\tau \propto 1 - w_\tau/\theta_\tau$ (wrt. scaling of the final result), and $\hat{\beta}_k = \hat{\beta}_\tau \gamma_k$ for $\rho_k \sqsubset_! \rho_\tau$, where*

$$\theta_\tau \triangleq \Big\| g_{\hat{\rho}_\tau} + \sum_{\rho_k \sqsubset_! \rho_\tau} \gamma_k g_{\hat{\rho}_k} \Big\|,$$

and γ_k are the multipliers for the smaller problem with the τ-component removed: $w_\tau = 0$ and $g_{\hat{\rho}_\tau} = 0$.

Proof. The problem (29) is a convex problem, and therefore the Karush-Kuhn-Tucker conditions being fulfilled is sufficient for a minimum. Let $\alpha_k \triangleq \|\rho_k z\| = \Big\| \sum_{\rho' \sqsubset \rho_k} \beta_{\rho'} g_{\rho'} \Big\|$. Then, inserting (30) into (29), differentiating wrt. β_ρ, adding the constraints $-\beta_\rho \leq 0$ and $\|z\|^2 \leq 1$, we get after dividing by $\|g_\rho\|^2$,

$$\begin{aligned} \lambda_\rho \geq 0,\ \lambda_\rho \beta_\rho = 0\, \forall \rho \in \psi, \quad \lambda \geq 0,\ \lambda(\|z\|^2 - 1) = 0 \\ 1 - \sum_{k\in\pi:\rho\sqsubset\rho_k} w_k \delta\Big(\frac{\beta_\rho}{\alpha_k}\Big) - \lambda\beta_\rho + \lambda_\rho \ni 0, \forall \rho \in \psi, \end{aligned} \tag{31}$$

where $\delta(\cdot)$ is a formal expression for handling non-differentiability. (If $\|g_\rho\|^2$ is zero, the condition for ρ may still be inserted, because the result does not then depend on β_ρ.) We may take $\lambda = 1$, for by positive homogeneity of h, the constraint on the norm is active unless the minimum is zero, and for any solution $\{\beta_\rho\}$ with $\lambda = \lambda' > 0$, $\{\lambda'\beta_\rho\}$ is a solution for $\lambda = 1$ (β_ρ/α_k being independent of such scaling).

For the maximal ρ_τ, by assumption $\alpha_\tau = \|\rho_\tau \hat{z}\| > 0$. Therefore $\hat{\beta}_\tau/\alpha_\tau > 0$ is defined, and (31) becomes for $\hat{\rho}_\tau$,

$$1 - w_\tau \frac{\hat{\beta}_\tau}{\alpha_\tau} - \hat{\beta}_\tau = 0,$$

so that $\hat{\beta}_\tau = \gamma_\tau \triangleq 1 - w_\tau/\theta'_\tau$ with $\theta'_\tau \triangleq \alpha_\tau/\hat{\beta}_\tau$. If $\gamma_\tau \leq 0$, our assumptions must be wrong, and $\hat{\beta}_\tau = 0$. So suppose this is not so.

If ρ_τ is also minimal, we get $\gamma_\tau = 1 - w_\tau/\|g_{\hat{\rho}_\tau}\|$, so that it is fully determined, and $\theta'_\tau = \theta_\tau$. Otherwise, set $\hat{\beta}_k = \gamma_k \hat{\beta}_\tau$ for some unknown γ_k for $\rho_k \sqsubset_! \rho_\tau$. Then also $\theta'_\tau = \theta_\tau$, and (31) becomes for ℓ with $\rho_\ell \sqsubset_! \rho_\tau$,

$$1 - w_\tau \frac{\gamma_\ell \hat{\beta}_\tau}{\alpha_\tau} - \sum_{k\in\pi:k\neq\tau,\hat{\rho}_\ell\sqsubset\rho_k} w_k \delta\Big(\frac{\gamma_\ell}{\alpha'_k}\Big) - \gamma_\ell \hat{\beta}_\tau - \lambda_{\hat{\rho}_\ell} \ni 0$$

where $\alpha'_k \triangleq \alpha_k/\hat{\beta}_\tau = \left\|\sum_{\hat{\rho}_j \sqsubset \rho_k} \gamma_j g_{\hat{\rho}_j}\right\|$. But $\gamma_\ell \hat{\beta}_\tau (1 + w_\tau/\alpha_\tau) = \gamma_\ell$, so that we get the condition

$$1 - \sum_{k \in \pi' : \hat{\rho}_\ell \sqsubset \rho_k} w_k \delta\Big(\frac{\gamma_\ell}{\alpha'_k}\Big) - \gamma_\ell - \lambda_{\hat{\rho}_\ell} \ni 0, \quad \forall \hat{\rho}_\ell \in \psi'$$

for $\pi' \triangleq \pi \setminus \{\tau\}$ and $\psi' \triangleq \psi \setminus \{\hat{\rho}_\tau\}$. This is a smaller problem of the original form. □

Note that the assumption $\|\rho_\tau \hat{z}\| > 0$ follows from $g_{\hat{\rho}_\tau} \neq 0$ by $\hat{\beta}_\tau > 0$. The lemma suggests the following method to find the multipliers $\hat{\beta}_k$: assume $\hat{\beta}_\tau > 0$ for maximal ρ_τ. Recursively repeat the procedure for the maximal $\rho_k \sqsubset_! \rho_\tau$ from the smaller problems defined by the lemma, until ρ_k is also minimal, in which case the lowest-depth factor $1 - w_k/\|g_{\hat{\rho}_k}\|$ can readily be calculated. Then calculate the higher factors $1 - w_\tau/\theta_\tau$ based on the information obtained from the deeper recursion levels. Finally scale the result. (This is not strictly necessary: the step size bounds $\alpha_0(\omega, \tilde{z}, v; p)$ include the scaling.) If ever $1 - w_\tau/\theta_\tau \leq 0$, the original assumption must be wrong, and we must have $\beta_\tau = 0$. This could result in a new set of problems, but we do actually have the following:

Theorem 8. *Lemma 12 continues to hold without the assumption $\hat{\beta}_\tau > 0$, so that we have (modulo scaling the final result) $\hat{\beta}_\tau = \max\{0, 1 - w_\tau/\theta_\tau\}$ for maximal ρ_τ, and $\hat{\beta}_k = \hat{\beta}_\tau \gamma_k$ for $\rho_k \sqsubset_! \rho_\tau$, with γ_k defined recursively from smaller problems.*

Proof. If $\hat{\beta}_\tau = 0$, and $\alpha_\tau > 0$, as we have assumed, then $\delta(\hat{\beta}_\tau/\alpha_\tau) = 0$, as there are no differentiability troubles. But then the condition (31) for maximal $\hat{\rho}_\tau$ becomes $1 + \lambda_{\hat{\rho}_\tau} = 0$, which has no solution, since $\lambda_{\hat{\rho}_\tau} \geq 0$. Therefore, the only way for $\hat{\beta}_\tau$ to be zero, is to have $\alpha_\tau = \|\rho_\tau z\| = 0$, so that $\delta(\hat{\beta}_\tau/\alpha_\tau)$ is not a singleton. But $\alpha_\tau = 0$ says that we can choose $\hat{\beta}_k = 0$ for all $\rho_k \sqsubset_! \rho_\tau$. □

Remark 6. Theorem 2 continues to hold with the range non-overlap assumption replaced by non-partial overlap assumption: If $z = -\beta_{\hat{\rho}_k} g_{\hat{\rho}_k} \neq 0$ for some $k \in \pi(q) \setminus \pi'$ with minimal ρ_k, then (since we have assumed $\rho_{\pi'} z = 0$), $0 > g^T z + \|z\|_k$ with $z \in \mathcal{R}(\hat{\rho}_k)$. The argument of Lemma 3 may therefore be applied to this subproblem to show deflection for k. Otherwise, k may be ignored (considered to be in π' for the purposes of this argument), and we may repeat the argument recursively.