JYVÄSKYLÄ STUDIES IN COMPUTING
99

Tuomo Valkonen

Diff-convex Combinations of Euclidean Distances

A Search for Optima

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN COMPUTING 99

Tuomo Valkonen

Diff-convex Combinations of Euclidean Distances

A Search for Optima

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksessa (Ag Aud. 2)
joulukuun 22. päivänä 2008 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in the Building Agora, Ag Aud. 2, on December 22, 2008 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2008

Diff-convex Combinations of Euclidean Distances

A Search for Optima

JYVÄSKYLÄ STUDIES IN COMPUTING 99

Tuomo Valkonen

Diff-convex Combinations of Euclidean Distances

A Search for Optima

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2008

Editors
Timo Männikkö
Department of Mathematical Information Technology, University of Jyväskylä
Pekka Olsbo, Marja-Leena Tynkkynen
Publishing Unit, University Library of Jyväskylä

URN:ISBN:978-951-39-7727-6
ISBN 978-951-39-7727-6 (PDF)

ISBN 978-951-39-3418-7 (nid.)
ISSN 1456-5390

Jyväskylä University Printing House, Jyväskylä 2008

ABSTRACT

Valkonen, Tuomo
Diff-convex combinations of Euclidean distances: a search for optima
Jyväskylä: University of Jyväskylä, 2008, 148 p.
(Jyväskylä Studies in Computing
ISSN 1456-5390; 99)
ISBN 978-951-39-3418-7
Finnish summary
Diss.

This work presents a study of optimisation problems involving differences of convex (diff-convex) functions of Euclidean distances. Results are provided in four themes: general theory of diff-convex functions, extensions of the Weiszfeld method, interior point methods, and applications to location problems.

Within the theme of general theory, new results on optimality conditions and sensitivity to perturbations of diff-convex functions are provided. Additionally, a characterisation of level-boundedness is provided, and the internal structure is studied for a class of diff-convex functions involving symmetric cones.

Based on this study, the Jordan-algebraic approach to interior point methods for linear programs on symmetric cones is extended. Local convergence of the method is proved, and a globalisation strategy is studied, based on the concept of the filter method.

The Weiszfeld method is extended to "perturbed spatial medians with incomplete data", where the convex spatial median objective function with scaled Euclidean distances can be perturbed by a concave function. The convergence of the method is studied, along with application to location problems.

The location problems of interest include in particular clustering and the Euclidean travelling salesperson problem (TSP). The classical multisource Weber problem is studied, and a new clustering objective is presented, based on a multiobjective interpretation of the problem. It is then shown that the Euclidean TSP can be presented as either of the above clustering objectives perturbed with a path length penalty.

The focus of the work is theoretical.

Keywords: Euclidean distance, diff-convexity, symmetric cone, interior point method, Weiszfeld method, clustering, travelling salesperson problem

Author

Tuomo Valkonen
Department of Mathematical Information Technology
University of Jyväskylä
Finland

Supervisors

Professor Tommi Kärkkäinen
Department of Mathematical Information Technology
University of Jyväskylä
Finland

Professor Marko Mäkelä
Department of Mathematics
University of Turku
Finland

Professor Pekka Neittaanmäki
Department of Mathematical Information Technology
University of Jyväskylä
Finland

Reviewers

Professor Vladimir F. Demyanov
Department of Applied Mathematics
St. Petersburg State University
Russia

Professor Marc Teboulle
School of Mathematical Sciences
Tel-Aviv University
Israel

Opponent

Professor Jean-Baptiste Hiriart-Urruty
Institut de Mathématiques
Université Paul Sabatier
Toulouse
France

ACKNOWLEDGEMENTS

To begin with, I would like to express my gratitude to my supervisors, professors Tommi Kärkkäinen, Marko Mäkelä, and Pekka Neittaanmäki. Their encouragement and support has greatly helped my research reach the stage shown on these pages.

A great deal of gratitude is as well due to professors Vladimir Demyanov and Marc Teboulle for reviewing the manuscript – not exactly a small amount of work. Other kind individuals have also read manuscripts of parts of the thesis along the course of the research. For this, I express my thanks to professors Dan Tiba and Eugene Stepanov.

For financial support, I have the COMAS graduate school and Agora Center to thank for.

Finally, I want to acknowledge the support of my parents and other close relatives. That support was of the utmost importance for my work.

Jyväskylä, November 2008
Tuomo Valkonen

CONTENTS

1 INTRODUCTION

The general theme of this thesis is the problem of finding the minimisers, or at least critical points, of functions that can be presented as the difference of two convex functions that, moreover, are combinations of (projected) Euclidean distances themselves – typically sums and maxima. Various *location problems* of considerable practical importance are representable in such a form. The simplest example is the convex problem of finding the *spatial median* of a set of points, but some *clustering* or *facility location* problems as well as the *Euclidean travelling salesperson problem* (TSP) are encompassed by this scheme. The study of these applications is one of the four sub-themes of the present thesis, introduced in Section 1.4, and further covered in the final chapters, 6 and 7.

Our focus is primarily theoretical, however, and the remaining sub-themes consist of analysis of algorithms for these problems, as well as some general mathematical results for differences of convex functions. This class of functions is introduced in Section 1.1 that follows, and some aspects are studied in Chapter 2. In Chapter 3 we further study the internal structure of differences of restricted support functions of slices of *symmetric cones*. Such functions, also briefly discussed in Section 1.2 below, include sums of Euclidean norms, of relevance to the general theme of this thesis.

Our first algorithmic theme is the extension of *interior point methods* for linear programs over symmetric cones, to the above class of functions. Our modelling and analysis of the methods and problem is based on the *Jordan algebraic* approach. In this context, we also encounter *filter methods* as a tool to globalise methods with only local convergence guarantees. This theme is further introduced in Section 1.2, and covered in detail in Chapter 4, based on the analysis of Chapter 3.

Our second algorithmic theme, and where the work on this thesis began, is the analysis of generalisations of the *Weiszfeld method*, conventionally for the above-mentioned problem of the spatial median, to more general problems, including incomplete data and concave perturbations to the objective function for the spatial median. This method is introduced in Section 1.3, analysed in Chapter 5, and applied in chapters 6 and 7.

Before the detailed coverage of each of these themes in the remaining chapters of this thesis, we now further introduce them in the following sections, and discuss the contributions of this thesis on a coarse level.

1.1 Diff-convexity

1.1.1 Convex functions

Recall from, e.g., the classic of Rockafellar [1972] that a function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is *convex* if it satisfies $f(\lambda y + (1-\lambda)y') \leq \lambda f(y) + (1-\lambda)f(y')$ for all $\lambda \in [0,1]$ and $y, y' \in \mathbb{R}^m$. It is called *proper* if $f(y) > -\infty$ and it is finite at some point. It is *closed* if the epigraph $\{(y,r) \mid r \geq f(y), y \in \mathbb{R}^m\}$ is closed, which for proper f is the same as lower-semicontinuity. In what follows, we will only consider proper closed convex functions.

Denoting the inner product of $\mathbb{R}^m$ by $\langle \cdot, \cdot \rangle$, we can define the (Fenchel) *subdifferential* $\partial f(y)$ of f at y as the set of *subgradients* z that satisfy

$$f(y') - f(y) \geq \langle z, y' - y \rangle \text{ for all } y' \in \mathbb{R}^m.$$

The directional derivative in the direction Δy is then given as $f'(y; \Delta y) = \sup\{\langle z, \Delta y \rangle \mid z \in \partial f(y)\}$. We often denote the range of the subdifferential by $\mathcal{R}(\partial f) \triangleq \bigcup_{y \in \mathbb{R}^m} \partial f(y)$.

Likewise, for $\epsilon \geq 0$, the approximate or ϵ-subdifferential $\partial_\epsilon f(y)$ is defined as the set of approximate subgradients z that satisfy

$$f(y') - f(y) \geq \langle z, y' - y \rangle - \epsilon \text{ for all } y' \in \mathbb{R}^m.$$

Then f will have its ϵ-minimum over $\mathbb{R}^m$ at y if and only if $0 \in \partial_\epsilon f(y)$.

The *convex conjugate* of f is defined as $f^*(z) \triangleq \sup_x \{\langle z, x \rangle - f(x)\}$. If f is proper and closed, $z \in \partial_\epsilon f(y)$ if and only if $y \in \partial_\epsilon f^*(z)$, and we have $f(y) + f^*(z) \leq \langle z, y \rangle + \epsilon$; see, e.g., the latter of the two volumes by Hiriart-Urruty and Lemaréchal [1993].

1.1.2 Differences of convex functions

We call the function g *diff-convex* (DC) if it can be represented as the difference of two convex functions f and v, denoted by $g(y) = f_v(y) \triangleq f(y) - v(y)$; introductions to the topic are provided by Hiriart-Urruty [1984],Horst and Thoai [1999], and Tuy [1995]. Such a representation is not unique, as can be seen by adding the same finite convex function to both f and v.

The class of diff-convex functions is important: many important problems have a natural diff-convex representation, as we will see even in this thesis. In fact, every twice continuously differentiable (C^2) function on $\mathbb{R}^m$ is diff-convex, and the set of diff-convex functions on a compact set of $\mathbb{R}^m$ is dense among the continuous functions on this set [Tuy, 1995].

Assuming f is proper and closed, and ν is finite-valued, the point y is a global minimiser of f_ν if and only if for all $\epsilon \geq 0$, $\partial_\epsilon \nu(y) \subset \partial_\epsilon f(y)$; see, e.g., Hiriart-Urruty [1988, 1995]. On the other hand, according to Dür [2003], if this condition holds for all $\epsilon \in [0, \overline{\epsilon}]$ for some $\overline{\epsilon} > 0$, then y is a local minimiser. This condition is however not necessary for local optimality, and in Chapter 2 of this thesis, we analyse additional requirements towards that end along with deriving related characterisations of strict optimality. The basic condition $\partial\nu(y) \subset \partial f(y)$ is, in any case, necessary but not sufficient for local optimality. In fact, a similar condition can be derived more generally by means of exhausters, discussed, for example, in a survey by Demyanov [2002].

In practise, checking all the inclusions in the above characterisations of optimality can be difficult, as can be finding points satisfying them, although there do exist approximation methods convergent to global optima. In fact, the minimisation of DC functions is generally NP-hard. Kearfott and Kreinovich [2005] have showed that this remains the case even for the subset of DC functions that contains all the convex functions, at least one non-convex function, and is closed under addition, multiplication by constants, and affine precomposition. This thesis also includes, in Chapter 7, a proof of NP-hardness of another subclass of DC functions, through transformation of the Euclidean TSP.

Given these difficulties, we often settle in this thesis, for what we call *ϵ-semi-criticality*. This we define to stand for $0 \in \partial_\epsilon^{\mathrm{DC}} f_\nu(y)$ ($\epsilon \geq 0$), where, following the sum rule for approximate subdifferentials of convex functions [cf. Hiriart-Urruty and Lemaréchal 1993], we have defined

$$\partial_\epsilon^{\mathrm{DC}} f_\nu(y) \triangleq \bigcup \{\partial_{\epsilon_1} f(y) - \partial_{\epsilon_2} \nu(y) \mid \epsilon_1 + \epsilon_2 = \epsilon,\ \epsilon_1, \epsilon_2 \geq 0\}. \tag{1.1}$$

Here the arithmetical difference of two sets is, as usual, defined as $A - B \triangleq \{x - y \mid x \in A, y \in B\}$. Note that when $\epsilon = 0$, the *semi-criticality* condition can be written as $\partial f(y) \cap \partial \nu(y) \neq \emptyset$.

Let ∂° denote the subdifferential of Clarke [1983], also covered in, e.g., Mäkelä and Neittaanmäki [1992]. We then have $\partial^\circ f_\nu(y) \subset \partial^\circ f(y) + \partial^\circ(-\nu)(y) = \partial f(y) - \partial \nu(y)$ with equality whenever either f or ν is differentiable by convexity and finiteness. Thus we see that semi-criticality is necessary for criticality in the sense $0 \in \partial^\circ f_\nu(y)$, and equivalent to it whenever either function is differentiable – almost everywhere in the interior of the common domain, by Rademacher's Theorem. Under the standing assumption of finite-valued ν, one may also easily observe the necessity of ϵ-semi-criticality for ϵ-minimality; cf., e.g., Section 4.4.1.

We note that (1.1) clearly depends on the choice of f and ν in the decomposition, and not just the difference $g = f_\nu$ itself: for example, let $\epsilon = 0$, and suppose that $\|\nabla f(0) - \nabla \nu(0)\| \in (0, 2)$. Then add to both f and ν the same function $y \mapsto \|y\|$ with $\partial\|0\| = \mathbb{B}(0, 1)$ and $\partial\|0\| - \partial\|0\| = \mathbb{B}(0, 2)$. Here and throughout this thesis, $\mathbb{B}(x, r)$ denotes the closed ball of radius r around x.

By writing $0 \in \partial_\epsilon^{\mathrm{DC}} f_\nu(y)$ as $z \in \partial_{\epsilon_1} f(y) \cap \partial_{\epsilon_2} \nu(y)$, we see by the above-mentioned convex subdifferential duality properties that the latter holds if and only if $y \in \partial_{\epsilon_1} f^*(z) \cap \partial_{\epsilon_2} \nu^*(z)$. Therefore there exist simultaneous "dual" solutions to $0 \in \partial_\epsilon^{\mathrm{DC}} f_\nu(y)$ and $0 \in \partial_\epsilon^{\mathrm{DC}} f_{\nu^*}^*(z)$, where $f_{\nu^*}^*(z) = f^*(z) - \nu^*(z)$. In

fact, there exists another important duality relationship related to this criticality duality. Assuming for simplicity that ν is finite-valued and closed, one may define the Fenchel conjugate f^*_ν as above for convex functions. Then $f^*_\nu(z) = \sup_{z' \in \operatorname{dom} \nu^*} \{f^*(z+z') - \nu^*(z')\}$ according to a result of Pshenichnyi also proved by Ellaia and Hiriart-Urruty [1986] and more generally Hiriart-Urruty [1986]. Setting $z = 0$ and equating with the definition, we therefore have $\inf_y f_\nu(y) + \sup_{z \in \operatorname{dom} \nu^*} f^*_{\nu^*}(z) = 0$. This provides an important duality relationship and optimality condition, which is exploited, e.g., in the DCA method of An and Tao [2005].

Finally, along with the characterisations of optimality in Chapter 2, we derive closely-related sensitivity and level-boundedness formulae. These properties can be important, respectively, in the study of behaviour of solutions to perturbed optimisation problems, and to ensure the boundedness of iterates in optimisation methods. While the latter generally follows from showing level-coercivity [see, e.g., Rockafellar and Wets, 1998], we provide relationships to the inclusion of $\mathcal{R}(\partial\nu)$ within $\mathcal{R}(\partial f)$, which is sometimes more easily checked. Also the "quality" of this inclusion, as $\mathcal{R}(\partial\nu) \subset \psi\mathcal{R}(\partial f)$ for some $\psi \in [0,1)$, plays a role in the complexity and convergence analyses of Chapter 4. In case of our reformulation of the Euclidean TSP in Chapter 7, through this property we are able to show that an ϵ-semi-critical point can be found in polynomial time.[1]

1.2 Interior point methods and Jordan algebras

1.2.1 Interior point methods

Interior point methods have their roots in Karmarkar's [1984] ground-breaking *potential-reduction* method for linear programming, as well as classical *barrier methods*, as considered by Fiacco and McCormick [1968]. Introductions to the linear and convex cases are provided by, e.g., Potra and Wright [2000], and to the general non-linear case by Forsgren et al. [2002]; we will merely sketch some overall ideas, and then move on to more specific cases of the present interest.

In application to constrained programming of the barrier function (or *path-following*) approach, the idea is to add to the objective function a weighted barrier function, defining a sequence of problems approximating the original problem. As the barrier function is chosen to approach infinity on the boundary of the region defined by the inequality constraints of the original problem, these approximate problems have interior solutions, and thus the constraints can be neglected. As the barrier weight decreases towards zero, then under second order conditions on the behaviour of the original objective function, the solutions to these modified problems will converge to a solution of the original problem, often along a continuous *central path* [Fiacco and McCormick 1968; Forsgren et al. 2002]. The

[1] In the sense of a *polynomial-time approximation scheme* [PTAS, see, e.g., Ausiello et al., 1999]; the dependency on ϵ is log-polynomial in $1/\epsilon$, quickly yielding high constant factors.

solution from a problem with higher parameter value, can presumably be used to help solving a problem with a lower parameter value, as in continuation methods.

In the potential-reduction approach, by contrast, the idea is to minimise a potential function, that in some sense includes the (barrier) weight decrease in the objective. Typically, after a suitable transformation, the objective function is linear in methods derived this way (while the constraints may be convex). However, the barrier function approach can also be used to yield similar methods in these special cases, and it is often not actually necessary to solve the sequence of problems corresponding to different weights near-exactly, or to follow the central path closely. Rather, at each iteration of the algorithm, it suffices to take both a *normal step* towards the central path, i.e., towards the solution of the problem corresponding to the present parameter value, as well as a *tangential step* with the intent of decreasing this parameter.

Consider the linear program

$$\min_p \langle c, p \rangle \quad \text{with} \quad Ap = b,\ p \in \mathbb{R}^m_+, \tag{1.2}$$

where $\mathbb{R}^m_+ \triangleq \{p \in \mathbb{R}^m \mid p \geq 0\}$ is the non-negative orthant of $\mathbb{R}^m$. By replacing the constraint $p \geq 0$ with the addition of the logarithmic barrier function $-\mu \sum_i \log p_i$ ($\mu > 0$) to the objective, one gets a series of problems with solutions tending towards the solutions of (1.2) under slight non-degeneracy assumptions. The Karush-Kuhn-Tucker (KKT) conditions for the barrier function problem turn out to be

$$Ap = b,\ A^*y + d = c,\ p \circ d = \mu e;\ p, d \in \mathbb{R}^m_+, \tag{1.3}$$

where we denote $p \circ d \triangleq (p_1 d_1, \ldots, p_m d_m)$, $e \triangleq (1, \ldots, 1) \in \mathbb{R}^m$, and (d, y) are variables for the dual problem $\max\{\langle b, y \rangle \mid A^*y + d = c,\ d \in \mathbb{R}^m_+\}$.

If one linearises (1.3) and hopes to reduce μ by a factor of $\sigma \in (0, 1)$, then provided $p, d \in \operatorname{int} \mathbb{R}^m_+$, one gets the linear system

$$A\Delta p = 0,\ A^*\Delta y + \Delta d = 0,\ p \circ \Delta d + d \circ \Delta p = \Delta q,$$

where $\Delta q \triangleq \sigma\mu e - p \circ d$ consists of the normal step $\mu e - p \circ d$ and tangential step $(\sigma - 1)\mu e$. The crude *primal-dual* method that follows, can be refined into a polynomial-time method for linear programming, as shown by Todd and Ye [1990] as well as Kojima et al. [1991] through a potential reduction analysis. These methods, and the generalisations to convex settings by Nesterov and Nemirovskii [1994] and Nesterov and Todd [1997], form the basis of the class of interior point methods considered in Chapter 4 of this thesis.

1.2.2 The Jordan-algebraic approach

A *Jordan algebra* $\mathcal{J}$ is basically a generalisation of many of the properties of the algebra of symmetric matrices on $\mathbb{R}^m$, when the product is defined as the symmetry-preserving $A \circ B = (AB + BA)/2$. In the general case, the product $\circ$ is assumed to be bilinear, commutative, and power-associative; for details

on the theory, we refer the reader to Faraut and Korányi [1994] or Koecher [1999], while Chapter 3 also contains a more detailed summary than the present one. A *Euclidean* Jordan algebra, possessing an associative inner product ($\langle x, y \circ z\rangle = \langle x \circ y, z\rangle$), will also have an unit element e. Its elements will have eigenvalues, wherefore also traces, determinants, non-integer powers, and various norms can be defined. The maximum number of distinct eigenvalues is called the *rank* of $\mathcal{J}$ and denoted by r in the sequel.

In fact, every finite-dimensional Euclidean Jordan algebra is a direct product of a small set of simple Euclidean Jordan algebras: those of quadratic forms on $\mathbb{R}^{m+1}$, the above-mentioned $m \times m$ real symmetric matrices, complex Hermitian $m \times m$ matrices, Hermitian $m \times m$ matrices with quaternion entries, and a special Albert algebra of 3×3 matrices with octonion entries.

Of particular importance is the cone $\mathcal{K}$ of positive-semidefinite elements of $\mathcal{J}$, and its interior of positive-definite elements, i.e., the set where all the eigenvalues are positive. This interior is a *symmetric cone*, i.e., a *self-dual* and *homogeneous* convex cone. The latter property means that the automorphisms of the cone, i.e., the invertible linear mappings Q such that $Q\mathcal{K} = \mathcal{K}$, act transitively on $\operatorname{int}\mathcal{K}$. That is, for every $x, y \in \operatorname{int}\mathcal{K}$, there is a Q such that $Qx = y$. This is important in relation to convergence-ensuring scaling transformations in optimisation methods. Indeed, symmetric cones are the same as the *self-scaled cones* of Nesterov and Todd [1997]. Furthermore, to each $w \in \operatorname{int}\mathcal{K}$, there corresponds a unique automorphism Q_w, the *quadratic representation* of w, which can be used to define local norms in $\mathcal{K}$. This is again useful in locally determining the interior of $\mathcal{K}$.

If we replace the cone $\mathbb{R}^m_+$ in (1.2) by the *cone of squares* $\mathcal{K}$ of a Jordan algebra $\mathcal{J}$, and apply the barrier $-\mu \log\det(p)$, we still get the equivalent of (1.3), with $\circ$ and e standing for the the corresponding operators and elements of $\mathcal{J}$. Various interior point algorithms also remain polynomial for the resulting problem [Faybusovich 1997b,a; Schmieta and Alizadeh 2001, 2003].

We are most interested in the Jordan algebra of quadratic forms. We write an element of the algebra as $p = (p^0, \bar{p})$, where $p^0 \in \mathbb{R}$ and $\bar{p} \in \mathbb{R}^m$. The product is defined as $p \circ d \triangleq (p^0 d^0 + \bar{p}^T \bar{d}, p^0 \bar{d} + d^0 \bar{p})$. This Jordan algebra has the important property that $\mathcal{K} = \{p \in \mathcal{J} \mid p^0 \geq \|\bar{p}\|\}$ is the *second-order* or *Lorentz cone*. It is of obvious importance in relation to optimisation with Euclidean norms. In particular, we can write $\|x\| = \max\{x^T \bar{p} \mid p^0 = 1, p \in \mathcal{K}\}$, which turns a nonlinear constraint on a norm, into linear and symmetric-cone constraints.

Thus we see that the KKT conditions for various sums of Euclidean norms, and in particular the extended Weber problem (1.4) below, can be reduced into the form of condition (1.3) with $\mu = 0$ [see, e.g., Andersen et al., 2000; Xue and Ye, 1997; Alizadeh and Goldfarb, 2003]. Higher values of μ then correspond to perturbation of these conditions to allow working within the interior of $\mathcal{K}$. In fact, as shown in Chapter 3, the conditions (1.3) then correspond to $0 \in \partial_{r\mu} f(y)$ along with $p \circ d = \mu e$ forcing a particular "selection" within an expanded substructure of $\partial_{r\mu} f$. Alternatively, the conditions (1.3) for $\mu > 0$ are obtained by smoothing $\|x\|$ by applying a barrier function in the above expansion.

In Chapter 4, we further extend this approach to diff-convex problems, em-

ploying ϵ-semi-criticality as defined using (1.1).

1.3 The Weiszfeld method

What is known as the Weiszfeld method, was first proposed by Weiszfeld [1937] for solving the *(Fermat-)Weber problem*, or the *spatial median* of a set of points in $\mathbb{R}^m$. That is, the problem in question is

$$\min_y \sum_{k=1}^{n} \|a_k - y\|,$$

where $\|\cdot\|$ is the Euclidean norm, and $a_1, \ldots, a_k \in \mathbb{R}^m$ are prescribed points. In this basic case, on the assumption $y \neq a_k$, the method itself is actually just a gradient descent method with a particular choice of step length, $1/\sum_k(1/\|a_k - y\|)$. The convergence on this assumption was proved by Kuhn [1973]. The method was extended and convergence proved for the $y = a_k$ case by Ostresh [1978]. Various generalisations of the method exist to ℓ^p distances [Üster and Love, 2000; Morris, 1981] and more abstract settings [Eckhardt, 1980; Puerto and Rodríguez-Chía, 1999, 2006].

The extension of the method to incomplete data was proposed by Kärkkäinen and Äyrämö [2004, 2005] and partial convergence shown in Valkonen [2006, 2008a]. The problem is to find a solution to

$$\min_y \sum_{k=1}^{n} \|W_k(y - a_k)\|, \tag{1.4}$$

where W_k are diagonal positive-semidefinite matrices modelling the importance and incompleteness of the data a_k. This extension, where the step lengths are typically (that is, when $W_k(y - a_k) \neq 0$) calculated coordinate-wise with the above formula, is no longer generally a gradient descent method. In fact, it is shown in Valkonen [2006] that such a generalisation would have worse convergence properties than the proposed one, which also may not converge unless the data is simple enough.

In Chapter 5 we further extend the algorithm to problems involving a concave perturbation $-\nu$ to (1.4), making the problem a diff-convex one. This extension bears similar theoretical convergence properties as the above extension, if we replace "minimiser" with "semi-critical point".

1.4 Applications

We already covered the spatial median, or the Weber problem, in the previous section on the Weiszfeld method. The first obvious diff-convex extension of this convex problem is called the *Weber problem with attraction and repulsion*. It includes

some repulsive components with negative weights in the sum. Although we have not studied the application of our methods to this problem, it is studied by, e.g., Chen et al. [1992] and Drezner and Wesolowsky [1991]. Another extension of the Weber problem is the multi-prototype version, considered below, among other clustering objectives. Other potential applications of (some of) our methods, include the *Euclidean TSP*, which is considered later in this section, and the *Euclidean Steiner tree problem*, which we briefly discuss in Appendix 4.

1.4.1 Clustering

The prototypical clustering objective is that of the the K-means and similar partitioning objectives based on different distances. We are most interested in the one that uses Euclidean distances: the *multisource Weber problem*, or *K-spatial-medians*. The typical formulation is

$$\min_{y_1,\ldots,y_K} \sum_{k=1}^{n} \min_{i=1,\ldots,K} \|a_k - y_i\|,$$

where the data $\{a_k\}$ are as before. Alternatively, the objective may be written with $\bar{y} = (y_1, \ldots, y_K)$ in the DC form

$$f(\bar{y}) - \nu_{\mathrm{KM}}(\bar{y}) \triangleq \sum_{k=1}^{n} \sum_{i=1}^{K} \|a_k - y_i\| - \sum_{k=1}^{n} \max_{j=1,\ldots,K} \sum_{i \neq j} \|a_k - y_i\|. \tag{1.5}$$

The standard method for this problem is the method developed by Cox [1957] for the related K-means problem, and consists of successively assigning vertices to clusters by closest prototype, solving the resulting K Weber problems, and repeating until there is no change in assignments. The convergence to critical points is proved by Selim and Ismail [1984], although the characterisation of local optimality in that paper is flawed. To this we include corrections in Appendix 1. The computational and statistical properties of the K-spatial-medians are studied extensively by Äyrämö [2006], whereas we study, in Chapter 6, the application of the perturbed Weiszfeld method to this latter formulation, mainly on the theoretical level.

The objective function (1.5) bears a multi-objective interpretation: f asks to place all cluster prototypes $y_1, \ldots, y_K$ as close as possible to the spatial median of the data, whereas $-\nu_{\mathrm{KM}}$ asks to place them as far as possible from the data belonging to other clusters. (See Miettinen [1999] for an introduction to multi-objective optimisation.) Such a multi-objective interpretation led us to propose an alternative clustering criteria, where we replace ν_{KM} with an objective that asks to place all the cluster centres as far as possible from each other,

$$\nu_{\mathrm{MO}} \triangleq \sum_{i<j} \|y_i - y_j\|.$$

The resulting problem, including the choice of suitable scalarisation parameters (weights for ν_{MO}), is studied primarily in Chapter 6 with the addition of a few notes in Chapter 4.

1.4.2 The Euclidean travelling salesperson problem

It turns out that, as shown in Chapter 7, both of the above clustering objectives can, with very small modifications, be turned into objectives for the Euclidean travelling salesperson problem. This is the problem of finding a permutation $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$, such that the length of the path traversing all the points $a_1, \ldots, a_n$, is minimised,

$$\min_{\sigma} \sum_{i=1}^{n} \|a_{\sigma i} - a_{\sigma(i+1)}\|,$$

identifying a_{n+1} with a_1.

Let us define $f_{\mathrm{TSP}}(\bar{y}) \triangleq \|y_i - y_{(i+1)}\|$ (with the same identification as above). Then, as shown in Chapter 7, the solutions to (1.5) with added λf_{TSP}, are the solutions of the Euclidean TSP for $\lambda \in (0, 1/2)$. Likewise, there exists a $\hat{\lambda} > 0$, such that minimisers of

$$f(\bar{y}) + \lambda f_{\mathrm{TSP}}(\bar{y}) - \nu_{\mathrm{MO}}(\bar{y})$$

related to the "MO" clustering objective, are solutions of the Euclidean TSP for $\lambda \in (0, \hat{\lambda})$. Since f_{TSP} for $\lambda \leq 1$ is subsumed into ν_{MO}, the perturbed Weiszfeld method is still applicable to this problem, although that is not the case for the "KM" formulation. The interior point methods of Chapter 4, however, are applicable to both formulations.

These problems are studied in further detail in Chapter 7, along with developing some heuristic approaches for performance improvements.

Chronology and publications

Chronologically, the contents of this thesis should be ordered as follows: the results in Chapters 5 and 6 along with Section 2.5, some improvements and corrections aside, were achieved in 2006 as a continuation of the paper Valkonen [2006, 2008a]. These results have been submitted as Valkonen and Kärkkäinen [2008a]. Chapter 7 is a further development of that research, as a study of application of the perturbed Weiszfeld method to the Euclidean TSP. That research was largely performed in 2006–2007, and has been published as Valkonen and Kärkkäinen [2008b]. After that, during the autumn 2007 and early 2008, a study of a potentially improved method was embarked on, encouraged by the performance of convex interior point methods. The results of that research are found in Chapters 3 and 4, and have been submitted as Valkonen [2008c]. Finally, the results of Chapter 2, excluding Section 2.5, were achieved during a short period of the spring 2008, as an offshoot of the preceding Introduction. They have been submitted as Valkonen [2008b].

2 SOME PROPERTIES OF DIFF-CONVEX FUNCTIONS

2.1 Introduction

Let f and ν be proper closed convex functions on $\mathbb{R}^m$, with ν finite-valued. We define the difference of these functions as $f_\nu \triangleq f - \nu$. As shown in particular by Hiriart-Urruty [1988, 1995], a necessary and sufficient condition for $\hat{y} \in \mathbb{R}^m$ to be a global minimiser of f_ν, is that

$$\partial_\epsilon \nu(\hat{y}) \subset \partial_\epsilon f(\hat{y}) \quad \text{for all } \epsilon \geq 0. \tag{2.1}$$

For local optimality, Dür [2003] has showed the sufficiency of the existence of $\overline{\epsilon} > 0$, such that (2.1) holds for all $\epsilon \in [0, \overline{\epsilon})$. This condition is, however, not necessary for local optimality. In this chapter, we show that necessity follows under the additional constraint of the set of "mutual linearity" of f and ν around $\hat{y}$, being the singleton $\{\hat{y}\}$.

We also show, that the condition on mutual linearity along with a strict inclusion in (2.1) for $\epsilon \in (0, \overline{\epsilon})$ – but importantly not necessarily for $\epsilon = 0$ – is both necessary *and* sufficient for strict local optimality. Also, when f_ν is level-bounded, it turns out that strict inclusion for all $\epsilon > 0$ and a singleton mutual linearity set, is both necessary and sufficient for the uniqueness of $\hat{y}$ as a global minimiser.

Also in this chapter, we provide some formulae for the sensitivity of minimisers, as the function f_ν is subject to perturbations. We are able to bound such minimisers in a scaled polar of a star-difference $\partial_\epsilon f(\hat{y}) \stackrel{*}{-} \partial_\epsilon \nu(\hat{y})$, guaranteed to be bounded by the strict optimality conditions. In our analysis, we apply and modify the epigraphical methods of Attouch and Wets [1993, 1991], also found and refined in Rockafellar and Wets [1998]. Finally, we study the relationship of level-boundedness to the inclusion $\mathcal{R}(\partial\nu) \subset \mathcal{R}(\partial f)$, which is a "limiting version" of the inclusions $\partial_\epsilon \nu(y) \subset \partial_\epsilon f(y)$ seen in the discussed characterisations of optimality.

The rest of this chapter is organised as follows. In Section 2.2 we introduce notation and concepts employed in the later analysis. In Section 2.3 we provide

the aforementioned characterisations of optimality. Section 2.4 concentrates on the sensitivity analysis, and we conclude the chapter with the level-boundedness analysis of Section 2.5.

2.2 Definitions

We denote the *support function* of a convex set A by $\sigma(x; A) \triangleq \sup\{\langle z, x\rangle \mid z \in A\}$, and the *gauge* by $\psi_A(x) \triangleq \inf\{t \geq 0 \mid x \in tA\}$. The *normal cone* is defined as $N_A(x) \triangleq \{z \in \mathbb{R}^m \mid \langle z, x' - x\rangle \leq 0 \text{ for all } x' \in A\}$, and the *polar* by $A^\circ \triangleq \{z \mid \langle z, x\rangle \leq 1 \text{ for all } x \in A\}$. The *star-difference* is defined for two sets A and B as

$$A \stackrel{*}{-} B \triangleq \{z \mid z + B \subset A\}.$$

Note that this set is closed and convex, if both A and B are. The *closure*, *boundary*, *interior*, and *relative interior* of a set A are denoted, respectively, by $\operatorname{cl} A$, $\operatorname{bd} A$, $\operatorname{int} A$, and $\operatorname{ri} A$.

We say that a function $g : \mathbb{R}^m \to \mathbb{R}$ is *level-bounded*, if the *level sets* $\operatorname{lev}_c g \triangleq \{y \mid g(y) \leq c\}$ are bounded for all $c \in \mathbb{R}$.

We denote the *domain* of a convex function by $\operatorname{dom} f = \{y \mid f(y) < \infty\}$, which is non-empty for our functions of interest. We recall that the (Fenchel) *ϵ-subdifferential* of f at $y \in \mathbb{R}^m$ is defined as the set $\partial_\epsilon f(y)$ of $z \in \mathbb{R}^m$ that satisfy

$$f(y') - f(y) \geq \langle z, y' - y\rangle - \epsilon \quad \text{for all } y' \in \mathbb{R}^m$$

for a given $\epsilon \geq 0$. When $\epsilon = 0$, this definition reduces to the convex *subdifferential*, denoted by ∂f. We denote the *range* of the subdifferential by $\mathcal{R}(\partial f) \triangleq \bigcup_{y \in \mathbb{R}^m} \partial f(y)$. Our general reference for many of the basic properties of ϵ-subdifferentials listed below is provided by Hiriart-Urruty and Lemaréchal [1993].

Defining the convex graphs

$$G_f(y) \triangleq \operatorname{Graph}(\epsilon \mapsto \partial_\epsilon f(y)) = \{(z, \epsilon) \mid \epsilon \geq 0,\ z \in \partial_\epsilon f(y)\},$$

we have the expression

$$\begin{aligned} f(y+h) - f(y) &= \sup\{\langle h, z\rangle - \epsilon \mid \epsilon > 0,\ z \in \partial_\epsilon f(y)\} \\ &= \sup\{\sigma(h; \partial_\epsilon f(y)) - \epsilon \mid \epsilon > 0\} \\ &= \sigma((h, -1); G_f(y)). \end{aligned} \tag{2.2}$$

Let us also recall the definition of the linearisation error,

$$e_f(y'; y, z) \triangleq f(y') - f(y) - \langle z, y' - y\rangle,$$

and the subdifferential transportation formula: if $z \in \partial_\eta f(y)$, then $z \in \partial_\epsilon f(y')$ for $\epsilon \geq \eta + e_f(y'; y, z)$. Now we may define the region of *mutual linearity* around y as

$$L(y) \triangleq \{y' \mid z \in \partial f(y) \cap \partial v(y),\ e_f(y'; y, z) = e_v(y'; y, z) = 0\}.$$

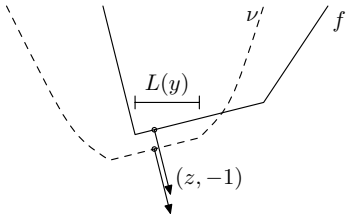


FIGURE 2.1 Illustration of the set $L(y)$. In this example $\partial f(y) = \partial \nu(y) = \{z\}$.

This region is illustrated in Figure 2.1.

Finally, we denote

$$C_\epsilon(y) \triangleq \partial_\epsilon f(y) \stackrel{*}{-} \partial_\epsilon \nu(y).$$

The condition $0 \in (\text{int})\ C_\epsilon(\hat{y})$ is then the same as $\partial_\epsilon \nu(\hat{y}) \subset (\text{int})\ \partial_\epsilon f(\hat{y})$, since $\partial_\epsilon \nu(\hat{y})$ is compact by our standing assumption on ν being finite-valued. Thus $0 \in \bigcap_{\epsilon>0} C_\epsilon(y)$ is equivalent to the necessary and sufficient global optimality condition $\partial_\epsilon \nu(y) \subset \partial_\epsilon f(y)$ for all $\epsilon > 0$. According to Martínez-Legaz and Seeger [1992], $\partial f_\nu(y) = \bigcap_{\epsilon>0} C_\epsilon(y)$, providing the connection to yet another characterisation of optimality.

2.3 Optimality

2.3.1 Strict local optimality

We may now state the main result of the present chapter.

Theorem 2.1. *The point $\hat{y} \in \mathbb{R}^m$ is a strict local minimiser of f_ν if and only if $L(\hat{y}) = \{\hat{y}\}$ and the following* subdifferential inclusion *is satisfied:*

$$\text{there exists } \overline{\epsilon} > 0, \text{ such that } 0 \in \operatorname{int} C_\epsilon(\hat{y}) \text{ for each } \epsilon \in (0, \overline{\epsilon}). \tag{SDI}$$

We begin the proof with a few lemmas.

Lemma 2.1. *Suppose $\hat{y} \in \operatorname{dom} f$, and that $(z_\nu, \epsilon_\nu) \in G_\nu(\hat{y}) \setminus \operatorname{int} G_f(\hat{y})$. Then there exists $(z_f, \epsilon_f) \in \operatorname{bd} G_f(\hat{y})$, $\alpha \geq 0$, and $(h, \delta) \in N_{G_f(\hat{y})}(z_f, \epsilon_f)$ with $\delta \in \{0, -1\}$ and $\|h\| \geq 1 + \delta$, such that $(z_\nu, \epsilon_\nu) = (z_f, \epsilon_f) + \alpha(h, \delta)$. We additionally have $\|h\| > 0$ if $\epsilon_\nu > 0$.*

Proof. Since $\hat{y} \in \operatorname{dom} f$, $G_f(\hat{y})$ is non-empty, in addition to being convex and closed. We may therefore choose $(z_f, \epsilon_f) \in G_f(\hat{y})$ as a constrained (not necessarily unique) minimiser of the function $(z, \epsilon) \mapsto \|(z_\nu, \epsilon_\nu) - (z, \epsilon)\|^2/2$, satisfying [see, e.g., Rockafellar, 1972, Theorem 27.4]

$$(z_\nu, \epsilon_\nu) - (z_f, \epsilon_f) \in N_{G_f(\hat{y})}(z_f, \epsilon_f). \tag{2.3}$$

We also have $(z_f, \epsilon_f) \in \operatorname{bd} G_f(\hat{y})$, because it was assumed that $(z_\nu, \epsilon_\nu) \in G_\nu(\hat{y}) \setminus \operatorname{int} G_f(\hat{y})$.

Because $\epsilon \mapsto \partial_\epsilon f(\hat{y})$ forms an increasing sequence of sets, we must have

$$(h, \delta) \in N_{G_f(\hat{y})}(z_f, \epsilon_f) \quad \text{implies} \quad \delta \leq 0. \tag{2.4}$$

Applying this to (2.3), we find that $\epsilon_f \geq \epsilon_\nu$.

If $\epsilon_f - \epsilon_\nu > 0$, we may set $\alpha \triangleq \epsilon_f - \epsilon_\nu$, and find $(h, -1) \in N_{G_f(\hat{y})}(z_f, \epsilon_f)$ after dividing (2.3) by α.

If $\epsilon_f = \epsilon_\nu$, there are two cases to consider. Suppose first that $z_f = z_\nu$. If there exists some $(h, -1) \in N_{G_f(\hat{y})}(z_f, \epsilon_f)$, we make this choice. Otherwise, we choose arbitrary $(h, 0) \in N_{G_f(\hat{y})}(z_f, \epsilon_f)$ with $\|h\| = 1$. Such a selection is guaranteed to exist by the observation (2.4), as well as the normal cone being non-zero at the boundary [see, e.g., Rockafellar, 1972, Corollary 11.6.1]. In both cases we set $\alpha = 0$.

If $z_f \neq z_\nu$, we choose $h = (z_\nu - z_f)/\alpha$ and $\delta = 0$ with $\alpha = \|z_\nu - z_f\|$.

Finally, if $\epsilon_\nu > 0$, recalling that also $\epsilon_f > 0$, it follows from $\partial_\epsilon f(\hat{y})$ being non-empty for $\epsilon \in (0, \epsilon_f)$, that $(0, -1) \notin N_{G_f(\hat{y})}(z_f, \epsilon_f)$. Therefore $\|h\| > 0$. □

Lemma 2.2. *Under the results of the preceding lemma, let $y_\lambda \triangleq \hat{y} + \lambda h$. Then, when $\delta = -1$, $f_\nu(y_\lambda) \leq f_\nu(\hat{y}) + (1-\lambda)\epsilon_\nu - \lambda\alpha$ for $\lambda \in [0, 1]$, and $z_f \in \partial f(y_1)$. In the case $\delta = 0$, $f_\nu(y_\lambda) \leq f_\nu(\hat{y}) + \epsilon_\nu - \lambda\alpha$ for all $\lambda \geq 0$.*

Proof. In all cases, applying $(z_\nu, \epsilon_\nu) = (z_f, \epsilon_f) + \alpha(h, \delta)$, we have

$$\begin{aligned} \nu(y_\lambda) - \nu(\hat{y}) &\geq \lambda\langle z_\nu, h\rangle - \epsilon_\nu = \lambda(\langle z_\nu, h\rangle - \epsilon_\nu) - (1-\lambda)\epsilon_\nu \\ &= \lambda(\langle z_f, h\rangle - \epsilon_f + \alpha(\|h\|^2 - \delta)) - (1-\lambda)\epsilon_\nu \\ &\geq \lambda(\langle z_f, h\rangle - \epsilon_f) + \lambda\alpha - (1-\lambda)\epsilon_\nu, \end{aligned} \tag{2.5}$$

where the last inequality follows from $\|h\| \geq 1 + \delta \in \{0, 1\}$.

Consider the $\delta = -1$ case first. By the expression (2.2), and the property $(h, -1) \in N_{G_f(\hat{y})}(z_f, \epsilon_f)$, we have

$$f(y_1) - f(\hat{y}) = \langle z_f, h\rangle - \epsilon_f.$$

This implies $e_f(y_1; \hat{y}, z_f) = -\epsilon_f$, whence $z_f \in \partial f(y_1)$. Furthermore, by convexity

$$f(y_\lambda) - f(\hat{y}) \leq \lambda(f(y_1) - f(\hat{y})) = \lambda(\langle z_f, h\rangle - \epsilon_f) \quad \text{for} \quad \lambda \in [0, 1]. \tag{2.6}$$

Thus the inequalities (2.6) and (2.5) imply as claimed,

$$f_\nu(y_\lambda) - f_\nu(\hat{y}) \leq (1-\lambda)\epsilon_\nu - \lambda\alpha.$$

Now, if $\delta = 0$, since $(h, 0) \in N_{G_f(\hat{y})}(z_f, \epsilon_f)$ with $\|h\| \geq 1 + \delta > 0$, we find that z_f maximises $\langle z, h\rangle$ over all $G_f(\hat{y})$. Consequently, the supremum in (2.2) is reached by $\epsilon \leq \epsilon_f$. Therefore

$$\begin{aligned} f(y_\lambda) - f(\hat{y}) &= \sup\{\lambda\langle z, h\rangle - \epsilon \mid 0 < \epsilon \leq \epsilon_f, z \in \partial_\epsilon f(\hat{y})\} \\ &\leq \sup\{\lambda\langle z_f, h\rangle - \epsilon \mid 0 < \epsilon \leq \epsilon_f\} = \lambda\langle z_f, h\rangle. \end{aligned}$$

Combining this with (2.5) yields the claim, since $\delta = 0$ implies $\epsilon_f = \epsilon_\nu$. □

Lemma 2.3. *Suppose* (SDI) *holds for $\hat{y}$ (possibly without the interior restriction), $z \in \partial\nu(y)$ and $\overline{\epsilon} > \epsilon_\nu \triangleq e_\nu(\hat{y}; y, z)$. Then*

$$f_\nu(y) - f_\nu(\hat{y}) \geq \sup\{\sigma(y - \hat{y}; C_\epsilon(\hat{y})) - (\epsilon - \epsilon_\nu) \mid \epsilon \in [\epsilon_\nu, \overline{\epsilon})\}. \tag{2.7}$$

Proof. By the subdifferential transportation formula, $z \in \partial_{\epsilon_\nu}\nu(\hat{y})$ as well as

$$\nu(\hat{y}) - \nu(y) = -\langle y - \hat{y}, z\rangle + \epsilon_\nu = -\sigma(y - \hat{y}; \partial_{\epsilon_\nu}\nu(\hat{y})) + \epsilon_\nu, \tag{2.8}$$

the latter equality following from the definition of the subdifferential.

Since $\overline{\epsilon} > \epsilon_\nu$, we furthermore get

$$\begin{aligned} f(y) - f(\hat{y}) &= \sup\{\sigma(y - \hat{y}; \partial_\epsilon f(\hat{y})) - \epsilon \mid \epsilon > 0\} \\ &\geq \sup\{\sigma(y - \hat{y}; \partial_\epsilon f(\hat{y})) - \epsilon \mid \epsilon \in [\epsilon_\nu, \overline{\epsilon})\} \\ &\geq \sup\{\sigma(y - \hat{y}; C_\epsilon(\hat{y})) + \sigma(y - \hat{y}; \partial_\epsilon\nu(\hat{y})) - \epsilon \mid \epsilon \in [\epsilon_\nu, \overline{\epsilon})\} \\ &\geq \sup\{\sigma(y - \hat{y}; C_\epsilon(\hat{y})) + \sigma(y - \hat{y}; \partial_{\epsilon_\nu}\nu(\hat{y})) - \epsilon \mid \epsilon \in [\epsilon_\nu, \overline{\epsilon})\}. \end{aligned} \tag{2.9}$$

Combining (2.8) and (2.9), we get (2.7). □

Proof of Theorem 2.1. Necessity. We may assume that $\hat{y} \in \operatorname{dom} f$, since otherwise $\hat{y}$ can not minimise f_ν *strictly*, even locally. That we must have $L(\hat{y}) = \{\hat{y}\}$ is clear from the definition of the linearisation error. To prove the necessity of (SDI), we assume the contrary, i.e., that there exists a sequence $\epsilon_{\nu,[k]} \searrow 0$ ($k \to \infty$), such that $0 \notin \operatorname{int} C_{\epsilon_{\nu,[k]}}(\hat{y})$. Then by the compactness of $\partial_{\epsilon_{\nu,[k]}}\nu(\hat{y})$, there exists $z_{\nu,[k]}$ with $(z_{\nu,[k]}, \epsilon_{\nu,[k]}) \in G_\nu(\hat{y}) \setminus \operatorname{int} G_f(\hat{y})$. Consequently Lemma 2.1 provides $(z_{f,[k]}, \epsilon_{f,[k]}) \in \operatorname{bd} G_f(\hat{y})$ and $(h_{[k]}, \delta_{[k]}) \in N_{G_f(\hat{y})}(z_{f,[k]}, \epsilon_{f,[k]})$ with $\delta_{[k]} \in \{0, -1\}$ and $\|h_{[k]}\| \geq 1 + \delta_{[k]}$, as well as $\alpha_{[k]} \geq 0$ such that $(z_{\nu,[k]}, \epsilon_{\nu,[k]}) = (z_{f,[k]}, \epsilon_{f,[k]}) + \alpha_{[k]}(h_{[k]}, \delta_{[k]})$.

First, consider the case that (for a subsequence) $\|h_{[k]}\| \to 0$. We may also assume that $\delta_{[k]} = -1$, as this must eventually be the case. Consequently, for $y_{[k]} \triangleq \hat{y} + h_{[k]}$, we have from Lemma 2.2 that $f_\nu(y_{[k]}) \leq f_\nu(\hat{y})$. But $y_{[k]} \to \hat{y}$, which provides a contradiction. We may therefore assume that $\|h_{[k]}\| \geq \theta > 0$.

The sequence $z_{\nu,[k]}$ is bounded by the finiteness of ν, and therefore may be assumed convergent to some $z_\nu \in \partial\nu(\hat{y})$, as $\epsilon_{\nu,[k]} \searrow 0$. It follows by construction from the boundedness of $(z_{\nu,[k]}, \epsilon_{\nu,[k]})$, that the sequence $(z_{f,[k]}, \epsilon_{f,[k]})$ is also bounded, and may likewise be assumed convergent to some $(z_f, \epsilon_f) \in \operatorname{bd} G_f(\hat{y})$. Since these considerations force $\alpha_{[k]}(h_{[k]}, \delta_{[k]})$ to be convergent, we may find $\alpha \geq 0$, and $(h, \delta) \in N_{G_f(\hat{y})}(z_f, \epsilon_f)$ such that[1] $\delta \in \{0, -1\}$, $\|h\| \geq \max\{1 + \delta, \theta\} > 0$ and $(z_\nu, 0) = (z_f, \epsilon_f) + \alpha(h, \delta)$.

The data at the limit therefore satisfies the assumptions of Lemma 2.2, and consequently, since $\epsilon_\nu = 0$, in either of the cases $\delta \in \{0, -1\}$, we have for $\lambda \in$

[1] The data (α, δ, h) cannot be chosen as a limit of a subsequence of $(\alpha_{[k]}, \delta_{[k]}, h_{[k]})$ only if $h_{[k]}$ contains no bounded subsequence, in which case $\alpha_{[k]} \to 0$. But then by the outer semicontinuity of $N_{G_f(\hat{y})}$, we may take h as a normalised limit of $h_{[k]}$, $\delta = 0$, and choose α to satisfy the sum constraint.

$[0,1]$ that $f(y_\lambda) \leq f(\hat{y}) - \lambda\alpha \leq f(\hat{y})$. Since $\|h\| > 0$, letting $\lambda \searrow 0$ provides a contradiction.

Sufficiency. Assume to the contrary, that there exists a sequence $y_{[k]} \to \hat{y}$ $(y_{[k]} \neq \hat{y})$, such that $f_\nu(y_{[k]}) \leq f_\nu(\hat{y})$. We may choose $z_{[k]} \in \partial\nu(y_{[k]})$ since $\operatorname{dom}\nu = \mathbb{R}^m$.

Let then $\epsilon_{\nu,[k]} \triangleq e_\nu(\hat{y}; y_{[k]}, z_{[k]})$. For sufficiently large k, we have $\epsilon_{\nu,[k]} < \overline{\epsilon}$, since $\{z_{[k]}\}$ is bounded, ν is continuous, and $y_{[k]} \to \hat{y}$. Therefore Lemma 2.3 applies, and we get

$$f_\nu(y_{[k]}) - f_\nu(\hat{y}) \geq \sup\{\sigma(y_{[k]} - \hat{y}; C_\epsilon(\hat{y})) - (\epsilon - \epsilon_{\nu,[k]}) \mid \epsilon \in [\epsilon_{\nu,[k]}, \overline{\epsilon})\}. \quad (2.10)$$

If $\epsilon_{\nu,[k]} > 0$, then choosing $\epsilon = \epsilon_{\nu,[k]}$, we have $0 \in \operatorname{int} C_\epsilon(\hat{y})$, so that (2.10) yields $f_\nu(y_{[k]}) - f_\nu(\hat{y}) > 0$, which is the desired contradiction.

If $\epsilon_{\nu,[k]} = 0$, we have $z_{[k]} \in \partial\nu(\hat{y})$, and therefore also $z_{[k]} \in \partial f(\hat{y})$, since it follows from (SDI) that $0 \in C_0(\hat{y})$. Therefore, as $e_\nu(y_{[k]}; \hat{y}, z_{[k]}) = 0$ by (2.8), and $y_{[k]} \neq \hat{y}$ by assumption, the condition $L(\hat{y}) = \{\hat{y}\}$ forces $e_f(y_{[k]}; \hat{y}, z_{[k]}) > 0$. But $e_f(y_{[k]}; \hat{y}, z_{[k]}) - e_\nu(y_{[k]}; \hat{y}, z_{[k]}) > 0$ says that $f_\nu(y_{[k]}) > f_\nu(\hat{y})$. □

Remark 2.1.

(i) Note that our conditions ensure $0 \in C_0(\hat{y})$ by closedness of the subdifferentials, but we *do not require* $0 \in \operatorname{int} C_0(\hat{y})$, which in itself is sufficient for strict local optimality, as shown by, e.g., Penot [1998] in a more general setting.

(ii) We have assumed ν to be finite-valued. This is not strictly necessary: all that is really needed is that the subdifferentials are uniformly bounded around $\hat{y}$. Clearly this follows if $\hat{y} \in \operatorname{int}\operatorname{dom}\nu$. Elsewhere, for a well-defined decomposition either $f_\nu(\hat{y}) = -\infty$, or there are points close to $\hat{y}$ with this property, i.e., $\hat{y} \in \operatorname{bd}\operatorname{dom}\nu$. In the former case, $\hat{y}$ is a minimiser, and $\partial_\epsilon\nu(\hat{y}) = \emptyset$. Thus there is no problem. In the latter case, $\hat{y}$ is not a minimiser. But since $\operatorname{dom}\nu \cup \operatorname{dom} f = \mathbb{R}^m$ for well-defined decompositions, the cone $N_{\operatorname{dom}\nu}(\hat{y})$ contains no non-zero vector in common with $N_{\operatorname{dom} f}(\hat{y})$. Therefore the subdifferential inclusion cannot hold [cf. Rockafellar, 1972, Theorem 25.6].

(iii) The condition $L(\hat{y}) = \{\hat{y}\}$ follows if f (or ν) is strictly convex, for then the sets $y \mapsto \{y' \mid e_f(y'; y, z) = 0,\ z \in \partial f(y)\}$ are singletons. Indeed, we have the following corollary.

Corollary 2.1. *The diff-convex function $g : \mathbb{R}^m \to (-\infty, \infty]$ has a strict local minimum at $\hat{y} \in \mathbb{R}^m$, if and only if* (SDI) *holds for* every *decomposition $f_\nu = g$ (with ν finite).*

Proof. The necessity is obvious from Theorem 2.1, while the sufficiency follows from choosing a decomposition f_ν with $L(\hat{y}) = \{\hat{y}\}$. This can be done by taking an arbitrary decomposition and adding the function $y \mapsto \theta\|y - \hat{y}\|^2$ for arbitrary $\theta > 0$ to both f and ν, to form the functions f^θ and ν^θ. Then $\partial f^\theta(\hat{y}) = \partial f(\hat{y})$, whence for all $z \in \partial f^\theta(\hat{y})$ and $y \neq \hat{y}$,

$$f^\theta(y) - f^\theta(\hat{y}) = f(y) - f(\hat{y}) + \theta\|y - \hat{y}\| > f(y) - f(\hat{y}) \geq \langle z, y - \hat{y}\rangle.$$

This says that $e_{f^\theta}(y; \hat{y}, z) > 0$. Since $g = f^\theta - \nu^\theta$, and by assumption (SDI) holds, Theorem 2.1 proves strict local optimality. □

The next example demonstrates that the condition $L(\hat{y}) = \{\hat{y}\}$ cannot be omitted, that is, (SDI) is not sufficient alone.

Example 2.1. Define the real functions

$$f(y) \triangleq \begin{cases} 0, & y \in (-1,1), \\ |y| - 1, & \text{otherwise}, \end{cases} \quad \text{and} \quad \nu(y) \triangleq f(y/2).$$

Then clearly $y = 0$ is a non-strict global minimiser of f_ν. But $\partial_\epsilon \nu(0) = \partial_\epsilon f(0)/2$, wherefore (SDI) holds, although strict optimality does not.

2.3.2 Non-strict local optimality

We now consider necessary conditions for local optimality, improving the sufficiency analysis of Dür [2003].

Theorem 2.2. *For the point $\hat{y} \in \mathbb{R}^m$ to be a local minimiser of f_ν, it is sufficient that*

$$\text{there exists } \overline{\epsilon} > 0, \text{ such that } 0 \in C_\epsilon(\hat{y}) \text{ for each } \epsilon \in [0, \overline{\epsilon}). \tag{SDI′}$$

If $L(\hat{y}) = \{\hat{y}\}$, this condition is also necessary.

(We do not require $\hat{y} \in \operatorname{dom} f$ for necessity, because $L(\hat{y}) = \{\hat{y}\}$ forces this.)

Proof. As mentioned, sufficiency has been shown in Dür [2003], but the proof of sufficiency in Theorem 2.1 can also be directly adapted by assuming the existence of a sequence $y_{[k]} \to \hat{y}$ with $f_\nu(y_{[k]}) < f_\nu(\hat{y})$, and then choosing $\epsilon = \epsilon_{\nu,[k]}$ in (2.10).

Necessity likewise follows by adapting the proof of Theorem 2.1. In the present situation, to reach a contradiction, we take $z_{\nu,[k]} \notin \partial_{\epsilon_{\nu,[k]}} f(\hat{y})$, wherefore $z_{f,[k]} \neq z_{\nu,[k]}$ and $\alpha_{[k]} > 0$.

Thus, in the case $\|h_{[k]}\| \to 0$, we actually have $f(y_{[k]}) \leq f(\hat{y}) - \alpha_{[k]} < f(\hat{y})$ with $y_{[k]} \to \hat{y}$, which is a contradiction.

Likewise, in the case $\|h_{[k]}\| \geq \theta > 0$, after choosing cluster points as in the proof of Theorem 2.1, we have $f(y_\lambda) \leq f(\hat{y}) - \lambda\alpha$ for $\lambda \in [0,1]$, wherefore $\alpha > 0$ provides a contradiction.

But if $\alpha = 0$, we must have $(z_f, \epsilon_f) = (z_\nu, 0)$. Then, from the proof of Lemma 2.2, for either choice of δ,

$$f(y_\lambda) - f(\hat{y}) \leq \lambda \langle z_f, h \rangle \quad \text{for } \lambda \in [0,1]. \tag{2.11}$$

Since $z_f \in \partial f(\hat{y})$, recalling that $\epsilon_f = 0$, the above must actually hold as an equality. Consequently $e_f(y_\lambda; \hat{y}, z_f) = 0$.

Now, if (2.5) holds strictly, i.e., $\nu(y_\lambda) - \nu(\hat{y}) > \lambda \langle z_f, h \rangle$, by combining with (2.11), we find a contradiction to local optimality as $\lambda \searrow 0$. But if we have equality for small λ, this means that $e_\nu(y_\lambda; \hat{y}, z_f) = 0$ with $z_f \in \partial\nu(\hat{y})$ (since $z_f = z_\nu$). Therefore $y_\lambda \in L(\hat{y})$. We have our contradiction, since $\|h\| > 0$ implies $y_\lambda \neq \hat{y}$. □

Similarly and with a proof analogous to Corollary 2.1, we get the following result. Notice the contrast between "every" and "some".

Corollary 2.2. *The diff-convex function $g : \mathbb{R}^m \to (-\infty, \infty]$ has a local minimum at $\hat{y} \in \mathbb{R}^m$, if and only if* (SDI′) *holds for* some *decomposition $f_\nu = g$ (with ν finite).*

Example 2.2. Dür [2003] provides a counterexample to the necessity of the existence of $\overline{\epsilon}$ without the additional assumption $L(\hat{y}) = \{\hat{y}\}$, in the form of

$$f(y) \triangleq \begin{cases} 0, & y \leq 1, \\ (y-1)^2, & y > 1, \end{cases} \quad \text{and} \quad \nu(y) \triangleq f(-y).$$

The function f_ν has local minimum at $y = 0$, but

$$\partial_\epsilon f(0) = [0, 2(\sqrt{1+\epsilon} - 1)], \quad \text{and} \quad \partial_\epsilon \nu(0) = -\partial_\epsilon f(0),$$

whence the condition $\partial_\epsilon \nu(0) \subset \partial_\epsilon f(0)$ does not hold for any $\epsilon > 0$. But we have $\{y \mid e_f(y; 0, 0)\} = (-\infty, 1]$, and $\{y \mid e_\nu(y; 0, 0)\} = [-1, \infty)$, whence $L(0) = [-1, 1]$.

2.3.3 Uniqueness of global minimisers

Finally, we represent some results pertaining to global optimality.

Theorem 2.3. *For the point $\hat{y}$ to be the unique global minimiser of f_ν, it is sufficient that $L(\hat{y}) = \{\hat{y}\}$ and* (SDI) *holds with $\overline{\epsilon} = +\infty$. If f_ν is level-bounded, this is also necessary.*

Proof. Only necessity demands further proof, sufficiency following from Lemma 2.3 completely analogously to the proof of Theorem 2.1 (without taking $y_{[k]} \to \hat{y}$).

Clearly again $L(\hat{y}) = \{\hat{y}\}$ is necessary, because f_ν takes on a single value on this set. Suppose that (z_ν, ϵ_ν) violates (SDI), with $\epsilon_\nu > 0$. As before, we apply Lemma 2.1 to the pair. Note that the resulting h is non-zero.

Now, if we have $\alpha > 0$ or $\delta = -1$, then Lemma 2.2 for suitable choice of λ provides a contradiction to uniqueness of the minimiser.

But if $\delta = 0$, by Lemma 2.2 $f_\nu(y_\lambda) \leq f(\hat{y}) + \epsilon_\nu$ on the set $\{y_\lambda \mid \lambda \geq 0\}$, which is unbounded since $h \neq 0$. This is in contradiction to level-boundedness. □

At this point, it is interesting to take a sneak peek into Corollary 2.6 to follow in Section 2.5. According to it, the inclusion $\operatorname{cl} \mathcal{R}(\partial \nu) \subset \operatorname{int} \mathcal{R}(\partial f)$ ensures level-boundedness, provided $\mathcal{R}(\partial \nu)$ is bounded. Thus, if the subdifferentials of ν are bounded and the interior inclusion condition (SDI) holds "in the limit", it is necessary that it holds for all $\epsilon > 0$, for $\hat{y}$ to be the unique global minimiser.

The following example shows that the strict inclusion condition does not necessarily hold without the additional level-boundedness assumption in Theorem 2.3.

Example 2.3. Consider

$$f(y) \triangleq \|y\|, \quad \text{and} \quad \nu(y) \triangleq \begin{cases} \|y\|^2/4, & \|y\| \leq 2, \\ \|y\| - 1, & \|y\| > 2. \end{cases}$$

Clearly $f_\nu(y) = 1$ outside $\mathbb{B}(0,2)$, while it has its unique global minimiser at $y = 0$. But (SDI) does not hold for $\epsilon \geq 1$, because

$$\partial_\epsilon f(0) = \mathbb{B}(0,1), \quad \text{and} \quad \partial_\epsilon \nu(0) = \begin{cases} \mathbb{B}(0,\sqrt{\epsilon}), & \epsilon \leq 1, \\ \mathbb{B}(0,1), & \epsilon > 1. \end{cases}$$

Remark 2.2. Theorem 2.3 could be refined. In the final case of the proof, since $\alpha = 0$, we have $(z_f, \epsilon_f) = (z_\nu, \epsilon_\nu)$. Therefore, since also $\delta = 0$, the procedure of Lemma 2.1 guarantees that there does not exist $(h', -1) \in N_{G_f(\hat{y})}(z_f, \epsilon_f)$. Thus it is merely necessary to have

$$z \in \partial_\epsilon f(\hat{y}) \cap \partial_\epsilon \nu(\hat{y}) \implies (h, -1) \notin N_{G_f(\hat{y})}(z, \epsilon) \text{ for all } h \in \mathbb{R}^m \tag{2.12}$$

along with $L(\hat{y}) = \{\hat{y}\}$ and (SDI′) for $\overline{\epsilon} = +\infty$.

We now show that this relaxed condition is sufficient as well: If y, z, and ϵ_ν are as in Lemma 2.3, they satisfy the premises of (2.12) by (SDI′). Because ri $G_f(\hat{y})$ is non-empty, the optimality characterisation [Rockafellar, 1972, Theorem 27.4] and (2.12) imply that $\langle (y - \hat{y}, -1), (z, \epsilon_\nu) \rangle$ cannot reach $\sigma((y - \hat{y}, -1); G_f(\hat{y}))$. We therefore have $f(y) - f(\hat{y}) > \langle z, y - \hat{y} \rangle - \epsilon_\nu$. Combining this estimate with (2.8), we get sufficiency.

2.4 Sensitivity

2.4.1 Local bounds for the inverse

Suppose $\hat{y}$ is a local minimiser of f_ν, satisfying (SDI′) (which necessarily follows in case of a strict minimiser). Let y' be another point, for which we have

$$\overline{\epsilon} > \epsilon_\nu \triangleq \min\{e_\nu(\hat{y}; y', z) \mid z \in \partial\nu(y')\}, \tag{2.13}$$

as well as the estimate

$$\eta \geq f_\nu(y') - f_\nu(\hat{y}).$$

Then, recalling Lemma 2.3, we have

$$\eta \geq \sup\{\sigma(y' - \hat{y}; C_\epsilon(\hat{y})) - (\epsilon - \epsilon_\nu) \mid \epsilon \in [\epsilon_\nu, \overline{\epsilon})\}. \tag{2.14}$$

But, since the set $C_\epsilon(\hat{y})$ is closed and convex, and contains the origin, the support function of this set is the gauge of the polar, $\sigma(\cdot; C_\epsilon(\hat{y})) = \psi_{C_\epsilon^\circ(\hat{y})}$, and the polar is closed and contains the origin [Rockafellar, 1972, Theorem 14.5]. Thus

$$\eta + (\epsilon - \epsilon_\nu) \geq \psi_{C_\epsilon^\circ(\hat{y})}(y' - \hat{y}) = \inf\{t \geq 0 \mid y' - \hat{y} \in tC_\epsilon^\circ(\hat{y})\}. \tag{2.15}$$

This says that (2.14) is equivalent to

$$y' \in \hat{y} + tC_\epsilon^\circ(\hat{y}) \quad \text{for all } t > \eta + \epsilon - \epsilon_\nu \text{ and } \epsilon \in [\epsilon_\nu, \overline{\epsilon}), \tag{2.16}$$

and then

$$y' \in \hat{y} + \bigcap_{\epsilon\in[\epsilon_\nu,\bar{\epsilon})} \bigcap_{t>\eta+\epsilon-\epsilon_\nu} tC_\epsilon^\circ(\hat{y}). \tag{2.17}$$

Now, if $t = \eta + \epsilon - \epsilon_\nu$ is not actually valid, the infimum is not reached in (2.15). But then it must actually be zero, because $C_\epsilon^\circ(\hat{y})$ is closed. Consequently, we may fix $t = \eta + \epsilon - \epsilon_\nu$ if this quantity is greater than zero. That can fail only if $\eta = 0$ and $\epsilon = \epsilon_\nu$, because $\eta \geq 0$ by (2.14) and $0 \in C_{\epsilon_\nu}(\hat{y})$. Therefore

$$y' \in \hat{y} + \bigcap_{\epsilon\in[\epsilon_\nu,\bar{\epsilon})} (\eta + \epsilon - \epsilon_\nu)C_\epsilon^\circ(\hat{y}) \qquad \text{when } \eta > 0. \tag{2.18}$$

In particular $y' \in \hat{y} + \eta C_{\epsilon_\nu}^\circ(\hat{y})$ for $\eta > 0$.

If $0 \in \operatorname{int} C_{\epsilon_\nu}(\hat{y})$, such as when $\hat{y}$ is a strict local minimiser (or the unique global minimiser of a level-bounded function) and $\bar{\epsilon}$ is chosen according to (SDI), then $C_{\epsilon_\nu}^\circ(\hat{y})$ is bounded. Consequently $t = 0$ yields no problem, and (2.18) holds. On the other hand, if $0 \in \operatorname{bd} C_{\epsilon_\nu}(\hat{y})$, then the polar is unbounded and we need to take the intersection over t.

If $\hat{y}$ is a global minimiser, we may take $\bar{\epsilon} = +\infty$, so the formula is valid for all points, while for a local minimiser, $\bar{\epsilon}$ and the linearisation error of ν bound the domain of applicability. Indeed, the formula (2.17) still depends on knowing ϵ_ν as defined in (2.13). Hence we still have the problem that our estimate is merely an a posteriori one. However, taking the union over ϵ_ν in (2.17), we get the following local restricted a priori estimate of the inverse of f_ν.

Lemma 2.4. *Suppose $\hat{y}$ is a local minimiser of f_ν, such that* (SDI′) *holds. Choose $\epsilon' \in [0, \bar{\epsilon})$, and let $D_{\hat{y}}(\epsilon') \triangleq \{y' \in \mathbb{R}^m \mid e_\nu(\hat{y}; y', z) \leq \epsilon',\ z \in \partial_\nu(y')\}$. When $y' \in D_{\hat{y}}(\epsilon')$ and $\eta \geq f_\nu(y') - f_\nu(\hat{y})$, we then have*

$$y' \in U_{\hat{y}}(\eta, \epsilon') \triangleq \hat{y} + \bigcup_{\epsilon_\nu\in[0,\epsilon']} \bigcap_{\epsilon\in[\epsilon_\nu,\bar{\epsilon})} \bigcap_{t>\eta+\epsilon-\epsilon_\nu} tC_\epsilon^\circ(\hat{y}).$$

When $\eta > 0$, $\epsilon > \epsilon_\nu$, or $0 \in \operatorname{int} C_{\epsilon_\nu}(\hat{y})$, we may fix $t = \eta + \epsilon - \epsilon_\nu$.

Corollary 2.3. *Let A be a non-empty set of global minimisers of f_ν, and choose $\epsilon' \geq 0$. Then $y' \in U_A(\eta, \epsilon') \triangleq \bigcup_{\hat{y}\in A} U_{\hat{y}}(\eta, \epsilon')$ whenever $y' \in D_A(\epsilon') \triangleq \bigcup_{\hat{y}\in A} D_{\hat{y}}(\epsilon')$ and $\eta \geq f_\nu(y') - \min f_\nu$.*

Remark 2.3. One could, of course, choose ϵ' independently for each $\hat{y}$, and extend the corollary to a set of local minimisers sharing the same function value.

Now, if y' is a (local) minimiser of a perturbed function, and we can approximate η for such points, we have a sensitivity result. Before calculating such bounds in Section 2.4.3 below, we look at some examples and study conditions ensuring the good behaviour of $U_{\hat{y}}(\eta, \epsilon')$ as $\eta \searrow 0$.

Example 2.4. Let $f(y) \triangleq \|y\|^2/2$, and $\nu(y) \triangleq \|y\|$. For simplicity we consider the situation on the real line, $m = 1$. Then the global minimisers of f_ν are $\hat{y} = \pm 1$. We fix $\hat{y} = 1$. Then

$$\partial_\epsilon f(1) = [1 - \sqrt{2\epsilon}, 1 + \sqrt{2\epsilon}], \quad \text{and} \quad \partial_\epsilon \nu(1) = [1 - \min\{\epsilon, 2\}, 1],$$

and consequently

$$C_\epsilon(1) = [\min\{\epsilon, 2\} - \sqrt{2\epsilon}, \sqrt{2\epsilon}].$$

Thus we may take $\overline{\epsilon} = +\infty$ in (SDI′). For the polars we get

$$C_\epsilon^\circ(1) = [1/(\min\{\epsilon, 2\} - \sqrt{2\epsilon}), 1/\sqrt{2\epsilon}].$$

Notice that $C_2^\circ(1) = [-\infty, 1/2]$, so the estimate becomes unbounded. But, in fact, $\epsilon_\nu = 0$ for $y' \in [0, \infty)$ and always $\epsilon_\nu \in \{0, 2\}$. Thus, after some tedious but elementary minimisation and maximisation calculations for an expression of $\bigcap_{\epsilon \geq 0}(\eta + \epsilon)C_\epsilon^\circ(1)$, we arrive at

$$y' \in 1 + \left[-\eta - \sqrt{\eta^2 + 2\eta}, \sqrt{2\eta}\right] \quad \text{when } \eta \in [0, 2] \text{ and } y' \in [0, \infty).$$

(For $\eta > 2$, optimal $\epsilon > 2$ in the lower bound, and the overall result would be a more complicated piecewise expression.) Now note that by symmetricity a mirror estimate holds around $\hat{y} = -1$, and the set $\{y' \mid e_\nu(\hat{y}; y', z) = 0, z \in \partial\nu(y'), \hat{y} = \pm 1\}$ covers the whole space. Therefore we have a well-behaved inverse estimate for the entire real line.

(The estimate remains bounded if we do indeed take the union of the estimates over $\epsilon_\nu \in [0, 2]$: While the exact expression is tedious to calculate, this can be seen by choosing, e.g., $\epsilon = 3$ for every ϵ_ν.)

Example 2.5. Exchanging the roles of f and ν in the previous example, we get that $\hat{y} = 0$ is a local minimiser, and that $C_\epsilon(0) = [\sqrt{2\epsilon} - 1, 1 - \sqrt{2\epsilon}]$ for $0 \leq \epsilon \leq \overline{\epsilon} \triangleq 1/2$. Therefore $C_\epsilon^\circ(0) = [1/(\sqrt{2\epsilon} - 1), 1/(1 - \sqrt{2\epsilon})]$. We may then approximate $\bigcap_{\epsilon \in [\epsilon_\nu, \overline{\epsilon})}(\eta + \epsilon - \epsilon_\nu)C_\epsilon^\circ(0) \subset \eta C_{\epsilon_\nu}^\circ(0)$. But then, choosing $\epsilon' \in (0, \overline{\epsilon}]$, we get $U_0(\eta, \epsilon') \subset \bigcup_{\epsilon_\nu \in [0, \epsilon']} \eta C_{\epsilon_\nu}^\circ(0) = \eta C_{\epsilon'}^\circ(0)$. Thus, for all y' with $e_\nu(y'; 0, 0) \leq \epsilon'$, i.e., $|y'| \leq \sqrt{2\epsilon'}$, we have the estimate $y' \in [\eta/(\sqrt{2\epsilon'} - 1), \eta/(1 - \sqrt{2\epsilon'})]$. The expression becomes unbounded as $\epsilon' \nearrow \overline{\epsilon} = 1/2$, i.e., as the region of validity closes $[-1, +1]$, the endpoints of which are global maxima, beyond which the function is decreasing to minus infinity; cf. the previous example.

Example 2.6. Let $\hat{y} \in \operatorname{dom} f$ be a local minimiser of a closed proper convex function f, and choose $\nu \equiv 0$. Then $\epsilon' = \epsilon_\nu = 0$, $\overline{\epsilon} = +\infty$, $D_{\hat{y}}(\epsilon') = \mathbb{R}^m$, and $C_\epsilon(\hat{y}) = \partial_\epsilon f(\hat{y})$. Assume $\eta > 0$ for simplicity. Then $y' \in U_{\hat{y}}(\eta, 0) = \hat{y} + \bigcap_{\epsilon \geq 0}(\eta + \epsilon)(\partial_\epsilon f(\hat{y}))^\circ$. In particular, $y' \in \hat{y} + \eta(\partial f(\hat{y}))^\circ$, which is of use if $0 \in \operatorname{int} \partial f(\hat{y})$.

Similar results continue to hold in a region of local convexity of f_ν, as exploited in Chapter 7 for sensitivity analysis of reformulations of the Euclidean TSP.

2.4.2 Continuity of the bounds

Recall that for a set-valued mapping F, the inner and outer limits are defined, respectively, as

$$\liminf_{x' \to x} F(x') \triangleq \{z \mid \text{ for all } x_{[k]} \to x \text{ there exist } F(x_{[k]}) \ni z_{[k]} \to z\}, \quad \text{and}$$
$$\limsup_{x' \to x} F(x') \triangleq \{z \mid \text{ there exist } x_{[k]} \to x \text{ and } F(x_{[k]}) \ni z_{[k]} \to z\}.$$

Clearly $\limsup F \supset \operatorname{cl} F(x) \supset \liminf F$. When both limits coincide, it is denoted $\lim F$, and when the common limit is $F(x)$, F is said to be continuous.

Lemma 2.5. *Suppose $\underline{\epsilon} \geq 0$. Then $\limsup_{\epsilon \searrow \underline{\epsilon}} C_\epsilon(\hat{y}) \subset C_{\underline{\epsilon}}(\hat{y})$. If* (SDI′) *holds with $\overline{\epsilon} > \underline{\epsilon}$ and either $C_{\underline{\epsilon}} = \{0\}$ or $\operatorname{int} C_{\underline{\epsilon}} \neq \emptyset$, then $\liminf_{\epsilon \searrow 0} C_\epsilon(\hat{y}) \supset C_{\underline{\epsilon}}(\hat{y})$. Consequently C_ϵ is continuous from above at $\underline{\epsilon}$.*

Proof. For the first inclusion, let $w_{[k]} \in C_{\epsilon_{[k]}}(\hat{y})$ converge to some w as $\epsilon_{[k]} \searrow \underline{\epsilon}$. Choose also $z_{[k]} \in \partial_{\epsilon_{[k]}} \nu(\hat{y})$ convergent to a given $z \in \partial_{\underline{\epsilon}} \nu(\hat{y}) = \bigcap_{\epsilon > \underline{\epsilon}} \partial_\epsilon \nu(\hat{y})$, recalling that this set is non-empty and bounded. Then also $\partial_{\epsilon_{[k]}} f(\hat{y}) \ni w_k + z_k \to w + z \in \partial_{\underline{\epsilon}} f(\hat{y})$. Since z was arbitrary, $w \in C_{\underline{\epsilon}}(\hat{y})$.

The second inclusion is immediate from (SDI′) in the case $C_{\underline{\epsilon}} = \{0\}$. In the case $\operatorname{int} C_{\underline{\epsilon}} \neq \emptyset$, we choose $\mu > 0$ small enough that $C^\mu \triangleq C_{\underline{\epsilon}} \stackrel{*}{-} \mathbb{B}(0, \mu) \neq \emptyset$. Next we choose $\epsilon > \underline{\epsilon}$ small enough that $\partial_\epsilon \nu(\hat{y}) \subset \partial_{\underline{\epsilon}} \nu(\hat{y}) + \mathbb{B}(0, \mu)$. Then

$$\partial_\epsilon \nu(\hat{y}) + C^\mu \subset \partial_{\underline{\epsilon}} \nu(\hat{y}) + C^\mu + \mathbb{B}(0, \mu) \subset \partial_{\underline{\epsilon}} \nu(\hat{y}) + C_{\underline{\epsilon}} \subset \partial_{\underline{\epsilon}} f(\hat{y}) \subset \partial_\epsilon f(\hat{y}).$$

That is, $C^\mu \subset C_\epsilon$ for small enough ϵ. But $\liminf_{\mu \searrow 0} C^\mu = C_{\underline{\epsilon}}$, because for all $z \in \operatorname{int} C_{\underline{\epsilon}}$, also $z \in C^\mu$ for small enough μ. □

Lemma 2.6. *Suppose* (SDI′) *holds, and that $\epsilon \mapsto C_\epsilon(\hat{y})$ is continuous from above at every $\epsilon \in [0, \epsilon')$. Then $U_{\hat{y}}(\eta_0, \epsilon') = \bigcap_{\eta > \eta_0} U(\eta, \epsilon')$ for all $\eta_0 \geq 0$.*

Proof. The inclusion $U_{\hat{y}}(\eta_0, \epsilon') \subset \bigcap_{\eta > \eta_0} U_{\hat{y}}(\eta, \epsilon')$ follows from the fact $\bigcap_i \bigcup_j x_{ij} \supset \bigcup_j \bigcap_i x_{ij}$.

For the other direction, suppose $y' \in U_{\hat{y}}(\eta, \epsilon')$ for all $\eta > \eta_0$. Reversing the argument that led to the definition of $U_{\hat{y}}$, we get that (2.14) holds for some $\epsilon_\nu = \epsilon_\nu(\eta) \in [0, \epsilon']$ and all $\eta > \eta_0$. Therefore, letting $\epsilon_\nu = \liminf \epsilon_\nu(\eta) \in [0, \epsilon']$ as $\eta \searrow \eta_0$, we have

$$\eta_0 \geq \sigma(y' - \hat{y}; C_\epsilon(\hat{y})) - (\epsilon - \epsilon_\nu) \quad \text{for all } \epsilon \in (\epsilon_\nu, \overline{\epsilon}). \tag{2.19}$$

Since $C_{\epsilon_\nu}(\hat{y}) \subset \liminf_{\epsilon \searrow \epsilon_\nu} C_\epsilon(\hat{y})$, the above holds for $\epsilon = \epsilon_\nu$ as well. Therefore (2.14) holds for $\eta = \eta_0$ and ϵ_ν. But this says $y' \in U_{\hat{y}}(\eta_0, \epsilon')$. □

For the next lemma, we directly extend the definition of the convex normal cone N_Q to possibly non-convex sets Q. Clearly, the property required below holds for N_Q if it holds for the regular normal cone $\widehat{N}_Q$; see Rockafellar and Wets [1998].

Lemma 2.7. *Suppose that $\hat{y}$ is a strict local minimiser of f_ν with $\overline{\epsilon} > 0$ satisfying* (SDI). *Let $Q \triangleq \operatorname{Graph}(\epsilon \mapsto C_\epsilon(\hat{y})) \cap \{\epsilon \leq \overline{\epsilon}\}$. Then $U_{\hat{y}}(0, \epsilon') = \{\hat{y}\}$ is equivalent to $(h, -1) \in N_Q(0)$ implying $h = 0$.*

Proof. The statement $y' \in U_{\hat{y}}(0, \epsilon')$ says that (2.14) holds for some $\epsilon_\nu \in [0, \epsilon']$ and $\eta = 0$. That is,

$$0 \geq \sup\{\sigma(y' - \hat{y}; C_\epsilon(\hat{y})) - (\epsilon - \epsilon_\nu) \mid \epsilon \in [\epsilon_\nu, \overline{\epsilon})\}. \tag{2.20}$$

If $\epsilon_\nu > 0$, choosing $\epsilon = \epsilon_\nu$ provides a contradiction, since $0 \in \operatorname{int} C_\epsilon(\hat{y})$. Thus $\epsilon_\nu = 0$ and therefore (2.20) says precisely that $(y' - \hat{y}, -1) \in N_Q(0)$. In consequence $y' = \hat{y}$, if the normal cone condition holds.

Conversely, $(y' - \hat{y}, -1) \in N_Q(0)$ implies (2.20) for $\epsilon_\nu = 0$. Then $y' \in U_{\hat{y}}(0, \epsilon')$. □

Theorem 2.4. *Under the conditions of Lemma 2.7 and the inner semi-continuity conditions of Lemma 2.5 for all $\underline{\epsilon} \in [0, \epsilon')$, $\bigcap_{\eta > 0} U_{\hat{y}}(\eta, \epsilon') = \{\hat{y}\}$.*

Proof. Apply the above three lemmas. □

The following example demonstrates the potential for failure of continuity in Lemma 2.5 when the dimension of $C_{\underline{\epsilon}}$ is not 0 or m.

Example 2.7. Let $\nu(y) \triangleq \|y\|^2/4$ and $f(y) \triangleq \max\{\nu(y), |y_2|\}$ when $y = (y_1, y_2)$. Then $\partial_\epsilon \nu(0) = \mathbb{B}(0, \sqrt{\epsilon})$, while [cf., e.g., Hiriart-Urruty and Lemaréchal, 1993, Theorem XI.3.5.1]

$$\begin{aligned}
\partial_\epsilon f(0) &= \bigcup\{\partial_{\epsilon_1}(\alpha_1 \nu)(0) + \partial_{\epsilon_2}(\alpha_2 |y_2|)(0) \mid {\scriptstyle \alpha_1+\alpha_2=1,\ \alpha_i \geq 0 \atop \epsilon_1+\epsilon_2 \leq \epsilon,\ \epsilon_i \geq 0}\} \\
&= \bigcup\{\alpha_1 \mathbb{B}(0, \sqrt{\epsilon_1/\alpha_1}) + \alpha_2(\{0\} \times \mathbb{B}(0,1)) \mid {\scriptstyle \alpha_1+\alpha_2=1,\ \alpha_i \geq 0 \atop \epsilon_1+\epsilon_2 \leq \epsilon,\ \epsilon_i \geq 0}\} \\
&= \bigcup\{\mathbb{B}(0, \sqrt{\alpha_1 \epsilon}) + \alpha_2(\{0\} \times \mathbb{B}(0,1)) \mid \alpha_1 + \alpha_2 = 1,\ \alpha_i \geq 0\}.
\end{aligned}$$

But now, because $\partial_\epsilon f(0)$ achieves the values $(\pm\sqrt{\epsilon}, y_2)$ for some y_2 only when $\epsilon = 0$ or $\alpha_1 = 1$, we find that $C_\epsilon = \{0\}$ for $\epsilon > 0$. On the other hand, $\partial \nu(0) = \{0\}$ and $\partial f(0) = \{0\} \times B(0,1)$, wherefore $C_0 = \{0\} \times B(0,1)$. Thus C_ϵ is not continuous from above at $\underline{\epsilon} = 0$.

The failure of the normal cone condition in Lemma 2.7 is demonstrated by the next example.

Example 2.8. Let $\nu(y) \triangleq y^2/4$ for $y \in \mathbb{R}$, and $f(y) \triangleq \nu(y)/(1 - |y|)$ when $|y| < 1$, and $+\infty$ otherwise. Clearly the function has a unique global minimum at $\hat{y} = 0$, so we set $\overline{\epsilon} = +\infty$. One can show through elementary manipulations that for $\epsilon \geq 0$, $\partial_\epsilon \nu(0) = \mathbb{B}(0, \sqrt{\epsilon})$, and $\partial_\epsilon f(0) = \mathbb{B}(0, \sqrt{\epsilon} + \epsilon)$. Thus $C_\epsilon(0) = [-\epsilon, \epsilon]$, so that in the notation of Lemma 2.7, Q is a self-dual cone, i.e., $N_Q(0) = -Q$. This violates the condition in the lemma. We also have $\bigcap_{\epsilon \in [\epsilon_\nu, \overline{\epsilon}]} (\eta + \epsilon - \epsilon_\nu) C_\epsilon^\circ(0) \supset [-1, 1]$ whenever $\epsilon_\nu < \eta$. Thus $\bigcap_{\eta > 0} U_0(\eta, \epsilon') \supset [-1, 1]$. Note that also $\nabla^2 f_\nu(0) = 0$.

2.4.3 The estimate η

We may apply the epigraphical methods of Attouch and Wets [1993, 1991], also covered in Rockafellar and Wets [1998], to finding an estimate η. However, a direct application of these findings results in sub-optimal results when we have a poor estimate of the epigraphical distance:

Let $\bar{\eta}$ be an "auxiliary epigraphical ρ-distance" of $g = f_\nu$ and another function $\tilde{g}$, defined as the the infimum of $\eta \geq 0$ that satisfy $\min_{\mathbb{B}(x,\eta)} \tilde{g} \leq \max\{g(x), -\rho\} + \eta$ and $\min_{\mathbb{B}(x,\eta)} g \leq \max\{\tilde{g}(x), -\rho\} + \eta$ for all $x \in \mathbb{B}(0,\rho)$. Suppose that ρ is sufficiently large (that it does not actually feature in the maxima), and that y' minimises $\tilde{g}$ in $\mathbb{B}(0,\rho)$. Then under some additional technical conditions, for small $\bar{\eta}$,

$$2\bar{\eta} \geq \min_{y \in \mathbb{B}(y',\bar{\eta})} (f_\nu(y) - \min f_\nu).$$

While we can deal with the minimisation, it and the factor two are unnecessary when we have poor maximum-difference estimates of the epigraphical distance. Consequently, we have the following result relying on a "two-sided distance". We denote $[D]_\delta \triangleq D \stackrel{*}{-} \mathbb{B}(0,\delta) = \{y \in D \mid \mathbb{B}(y,\delta) \in D\}$, $\gamma\text{-arg}\min g \triangleq \{x \mid g(x) \leq \min g + \gamma\}$, and by $\mathcal{F}$ the functions $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ that are lower-semicontinuous and level-bounded. (This ensures existence of minimisers).

Lemma 2.8. *Let $g, \tilde{g} \in \mathcal{F}$, and D be a closed set, such that $\min_D g < \infty$. Suppose $\delta, \tilde{\delta} \geq 0$ and $\eta, \tilde{\eta} \in \mathbb{R}$ are such that $A \triangleq \arg\min_D g \subset [D]_\delta$ and*

$$\min_{\mathbb{B}(y,\tilde{\delta})} g \leq \tilde{g}(y) + \tilde{\eta} \text{ for all } y \in D, \tag{2.21}$$

$$\min_{\mathbb{B}(y,\delta)} \tilde{g} \leq g(y) + \eta \text{ for all } y \in A. \tag{2.22}$$

Then $\eta + \tilde{\eta} + \gamma \geq \min_{\mathbb{B}(y',\tilde{\delta})}(g - \min_D g)$ whenever $y' \in \gamma\text{-arg}\min_D \tilde{g}$ ($\gamma \geq 0$).

Proof. Let $\hat{y} \in A$. Then by assumption $\mathbb{B}(\hat{y},\delta) \subset D$. Therefore, by (2.22), $\min_D \tilde{g} \leq \min_{\mathbb{B}(\hat{y},\delta)} \tilde{g} \leq \min_D g + \eta$.

Choose then $y' \in \gamma\text{-arg}\min_D \tilde{g}$. By (2.21) $\min_{\mathbb{B}(y',\tilde{\delta})} g \leq \tilde{g}(y') + \tilde{\eta} \leq \min_D \tilde{g} + \gamma + \tilde{\eta}$. Combine this with the result of the previous paragraph, to get the claim. □

Corollary 2.4. *Let $g, \tilde{g} \in \mathcal{F}$, and D be a closed set, such that $\min_D g < \infty$. Let $\eta \triangleq \sup_{D \cap \operatorname{dom} \tilde{g}}(g - \tilde{g}) - \inf_A(g - \tilde{g})$. Then $\eta + \gamma \geq g(y') - \min_D g$ whenever $y' \in \gamma\text{-arg}\min_D \tilde{g}$.*

Proof. Choose $\delta = \tilde{\delta} = 0$ in Lemma 2.8. □

2.4.4 The main sensitivity result

Let $f_\nu \in \mathcal{F}$ and $\emptyset \neq A \subset \arg\min f_\nu$. Choose $\epsilon' \geq 0$ as in Corollary 2.3, and suppose D is a closed set with $A \subset [D]_\delta$ and $D \subset [D_A(\epsilon')]_{\tilde{\delta}}$. Suppose that

$y' \in \gamma\text{-arg}\min_D \tilde{g}$. We then have $y \in D_A(\epsilon')$ for $y \in \mathbb{B}(y', \tilde{\delta})$. Therefore, for y achieving $\min_{\mathbb{B}(y',\tilde{\delta})}(f_\nu - \min_D f_\nu)$, we have $y \in U_A(\eta + \tilde{\eta} + \gamma, \epsilon')$ by Corollary 2.3 and Lemma 2.8. Thus, we have

Theorem 2.5. *Let $f_\nu = g, \tilde{g} \in \mathcal{F}$ and D be a closed set. Suppose $\delta, \tilde{\delta} \geq 0$ and $\eta, \tilde{\eta} \in \mathbb{R}$ are such that the assumptions of Lemma 2.8 hold along with $D \subset [D_A(\epsilon')]_{\tilde{\delta}}$. Then $y' \in U_A(\eta + \tilde{\eta} + \gamma, \epsilon') + \mathbb{B}(0, \tilde{\delta})$ whenever $y' \in \gamma\text{-arg}\min_D \tilde{g}$.*

Combining with Corollary 2.4, we get

Corollary 2.5. *Suppose $g = f_\nu, \tilde{g} \in \mathcal{F}$ and that D is a closed set with $A \subset D \subset D_A(\epsilon')$. Let $\eta \triangleq \sup_{D \cap \operatorname{dom} \tilde{g}}(g - \tilde{g}) - \inf_A(g - \tilde{g})$. Then $\gamma\text{-arg}\min_D \tilde{g} \subset U_A(\eta + \gamma, \epsilon')$.*

Remark 2.4. Instead of $D \subset [D_A(\epsilon')]_{\tilde{\delta}}$, we may assume $D \subset D_A(\epsilon')$ along with $\gamma\text{-arg}\min_D \tilde{g} \subset [D]_{\tilde{\delta}}$, and get similar results.

These results extend in a straightforward manner to sets of local minimisers A, provided that the minimum of f_ν on D is reached by all $\hat{y} \in A$. This extension is utilised in the following simple example.

Example 2.9. Consider the situation of Example 2.5. We have $D_0(\epsilon') = \sqrt{2\epsilon'} \cdot [-1, +1]$ and $U_0(\eta, \epsilon') \subset \eta/(1 - \sqrt{2\epsilon'}) \cdot [-1, +1]$ for $\epsilon' < 1/2$. Consider a simple tilted perturbation: $\tilde{g}(y) = f_\nu(y) + \lambda y$ for some $\lambda \in \mathbb{R}$. Then the function value does not change in $A = \{0\}$, and in $D_0(\epsilon')$, the maximum difference is $\eta = |\lambda|\sqrt{2\epsilon'}$. Thus we have $y' \in U_0(\eta, \epsilon') \subset |\lambda|\sqrt{2\epsilon'}/(1 - \sqrt{2\epsilon'}) \cdot [-1, +1]$ for $y' \in \arg\min_{D_0(\epsilon')} \tilde{g}$. Consequently, $y' \in \operatorname{int} D_0(\epsilon')$, i.e., y' is an unconstrained local minimiser, when $|\lambda| < 1 - \sqrt{2\epsilon'}$.

2.5 Level-boundedness

Theorem 2.6. *Suppose that f and ν are closed proper convex functions in $\mathbb{R}^m$, with $\mathcal{R}(\partial f)$ bounded. For the level sets $\operatorname{lev}_c f_\nu$ to be bounded, it is sufficient that $\operatorname{cl} \mathcal{R}(\partial \nu) \subset \operatorname{int} \mathcal{R}(\partial f)$ and necessary that $\mathcal{R}(\partial \nu) \subset \operatorname{int} \mathcal{R}(\partial f)$.*

From the assumption 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.

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

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 $z \in \operatorname{int} A$, we may rewrite $(f - \nu)(y) = (f(y) - \langle z, y\rangle) - (\nu(y) - \langle z, y\rangle)$, yielding another DC representation of the same function f_ν, for which $0 \in \mathcal{R}(\partial(f - \langle z, \cdot\rangle))$, and the required inclusion condition holds. Likewise we may assume that $\nu(0)$ is finite.

Denote $\tilde{v}^v_{\tilde{y}}(y) \triangleq \nu(\tilde{y}) + \langle v, y - \tilde{y}\rangle$. Since $\nu(y) = \sup_{\tilde{y}\in\mathbb{R}^m, v\in\partial\nu(\tilde{y})} \tilde{v}^v_{\tilde{y}}(y)$, with the supremum achieved (at least by $\tilde{y} = y$), we may expand

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

But, since $v \in \partial\nu(\tilde{y})$, we have $\nu(0) - \nu(\tilde{y}) \ge \langle v, 0 - \tilde{y}\rangle$, or that $\nu(\tilde{y}) - \langle v, \tilde{y}\rangle \le \nu(0)$. Hence, $\tilde{y}$ can be removed from the equation, and we have

$$\operatorname{lev}_c f_\nu \subset \bigcup_{v\in B} \{y \mid f(y) - \langle v, y\rangle \le c_0\} = \bigcup_{v\in B} \operatorname{lev}_{c_0}(f - v)$$

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 known from, e.g., Rockafellar [1966]. For the uniform boundedness of this family of sets, a little more work is needed.

By the inclusion $\operatorname{cl} B \subset \operatorname{int} A$, $0 \in A$ and $\operatorname{cl} A$ being convex [Rockafellar, 1972, Section 24] and bounded, every $v \in \operatorname{cl} B \setminus \{0\}$ has an $\epsilon_v \in (0, 1/4)$ such that $\mathbb{B}(v, 4\epsilon_v) \subset \operatorname{int} A$ and $v/(1 - 4\epsilon_v) \in \operatorname{cl} A$. Since $\operatorname{cl} B$ is a subset of the bounded set A, it is compact, and we can find a finite set $B^* \subset \operatorname{cl} B \setminus \{0\}$ such that the sets $v^* + 2\epsilon_{v^*}A$ for $v^* \in B^*$ cover B. 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 $v^* \in B^*$, which are finite in number.

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

$$|(f(y) - \langle v^*, y\rangle) - (f(y) - \langle v, y\rangle)| = |\langle v^* - v, y\rangle|.$$

But

$$\sup_{v\in v^*+2\epsilon_{v^*}A} |\langle v^* - v, y\rangle| = \sup_{z\in A} 2\epsilon_{v^*}|\langle z, y\rangle| = 2\epsilon_{v^*}|\langle z^*(y), y\rangle|$$

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

$$L_{v^*} \subset \{y \mid f(y) - \langle v^*, y\rangle \le c + 2\epsilon_{v^*}|\langle z^*(y), y\rangle|\}.$$

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

$$|\langle v^*, y\rangle| = (1 - 4\epsilon_{v^*})|\langle v^*/(1 - 4\epsilon_{v^*}), y\rangle| \le (1 - 4\epsilon_{v^*})|\langle z^*(y), y\rangle|,$$

and

$$L_{v^*} \subset \{y \mid f(y) \le c + (1 - 2\epsilon_{v^*})|\langle z^*(y), y\rangle|\}.$$

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

$$\begin{aligned}
f(y) &= \sup\{\langle z, y\rangle - f^*(z) \mid z \in \mathbb{R}^m\} \\
&\ge \sup\{\langle z, y\rangle - f^*(z) \mid z \in A'\} \\
&\ge (1 - \epsilon_{v^*})|\langle z^*(y), y\rangle| - \sup\{f^*(z) \mid z \in A'\}
\end{aligned}$$

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

$$L_{v^*} \subset \{y \mid \epsilon_{v^*} |\langle z^*(y), y\rangle| \leq c + c'\},$$

and this is clearly bounded, because we have assumed $0 \in \operatorname{int} A$, whence $|\langle z^*(y), y\rangle| \geq \delta \|y\|$ 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 int A [Rockafellar, 1972, Section 24]. Hence, if f^* was not bounded in A', a bounded set, we could find a sequence $\{z_{[k]}\}_{k=1}^{\infty} \subset A'$ converging to some $z \in \operatorname{bd} A'$ for which $f^*(z) = \infty$. But this contradicts the finiteness of f^* on int A.

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

Suppose then that $v \in \operatorname{bd} A \cap A$. Then $v \in \partial f(y)$ for some y. Let $h \in N_{\operatorname{cl} A}(v) \setminus \{0\}$. Then also $h \in N_{\partial_\epsilon f(y)}(v)$ for all $\epsilon \geq 0$. By (2.2), this says that for $y_\lambda \triangleq y + \lambda h$, $f(y_\lambda) = f(y) + \lambda\langle v, h\rangle$. Thus

$$\begin{aligned} f_\nu(y_\lambda) \leq f(y_\lambda) - \tilde{v}_{\tilde{y}}^{v}(y_\lambda) &= f(y_\lambda) - \nu(\tilde{y}) - \langle v, y_\lambda - \tilde{y}\rangle \\ &= f(y) - \nu(\tilde{y}) - \langle v, y - \tilde{y}\rangle, \end{aligned}$$

wherefore f_ν is bounded on the line $\lambda \mapsto y + \lambda h$. Therefore it has unbounded level sets. □

Example 2.10. To see that $\operatorname{cl} \mathcal{R}(\partial\nu) \subset \operatorname{int} \mathcal{R}(\partial f)$ is not necessary, consider the real functions $f : y \mapsto |y|$ and

$$\nu : y \mapsto \sup_{k=1,2,3,\ldots} \nu_k(y) \quad \text{with} \quad \nu_k(y) = \sum_{i=1}^{k} 2^{-i}(|y| - 2^i). \tag{2.23}$$

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

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

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

Example 2.11. 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 in (2.23).

Regarding conditions on f, we have the following extension:

Corollary 2.6. *The boundedness assumption on $\mathcal{R}(\partial f)$ in Theorem 2.6 can be lifted, provided $\mathcal{R}(\partial \nu)$ is bounded.*

Proof. The necessity proof does not depend on the boundedness assumption. As for sufficiency, 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 $\tilde{f}(y) \triangleq \sup\{f(\tilde{y}) + \langle z, y - \tilde{y}\rangle \mid z \in A \cap \partial f(\tilde{y})\}$. $\mathcal{R}(\partial \tilde{f})$ 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. □

Example 2.12. Similar conclusions do not necessarily follow if $\mathcal{R}(\partial \nu)$ is unbounded. This can be illustrated by considering the functions $y \mapsto \alpha y^2$ 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.

Lemma 2.9. *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 is 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 boundary 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 the entire $\mathbb{R}^m$. □

3 DIFF-CONVEX FUNCTIONS ON SYMMETRIC CONES

3.1 Introduction

In this chapter we consider functions $f_\nu = f - \nu$ expressible as the difference of convex functions of the form

$$f(y) \triangleq \sup\{\langle B^*y + c, p\rangle \mid p \in \mathcal{K}, Ap = b\}, \tag{3.1}$$

where $\mathcal{K}$ is (the closure of) a symmetric cone, c and b are constant vectors, and B and A are constant linear mappings such that the constraint set for p is non-empty and bounded. Our interest stems from potential applications and the fact that convex functions of the form (3.1) are important in relation to interior point methods. In particular, the necessary and sufficient optimality conditions for f may be written

$$B^*y + A^*\lambda + d + c = 0,\ Ap = b, Bp = 0,\ p \circ d = 0,\ p, d \in \mathcal{K}. \tag{3.2}$$

This condition is of the form (1.3), whence various efficient interior point methods are available for perturbed versions of (3.2) – which turn out to correspond to $0 \in \partial_\epsilon f(y)$.

In order to study the extension of these methods to f_ν in Chapter 4, we must analyse the solvability of the equivalent condition for f_ν. This is the main topic of the present chapter, covered in Section 3.4, and our tool is second order graphical differentiation. As additional consequences of our analysis we obtain an alternative derivation of the perturbed version of conditions (3.2) – often also derived through the use of barrier functions – as well as an alternative interpretation of what an "interior point" is. It could be said that this makes our approach in Chapter 4 "graphical programming".

First we, however, introduce in Section 3.2 some basic notations for the present and the following chapter, including a quick introduction to the Jordan-algebraic machinery used. The tangent and normal sets of ϵ-complementary pairs in a symmetric cone are also analysed in Section 3.3.

3.2 Preliminaries

3.2.1 Sets and mappings

First we introduce some basic notations. Let A be a mapping. Then $\mathcal{R}(A)$ denotes its range. When A is also linear, $\mathcal{N}(A)$ denotes its null-space. The adjoint of a linear operator A between two inner product spaces is denoted by A^*, and the pseudoinverse by $A^\dagger$. For two mappings, $(A, B)(x, y) \triangleq (Ax, By)$, and $(A; B)(x, y) \triangleq Ax + By$.

Let then C be a cone. Given an inner product $\langle \cdot, \cdot \rangle$, for the purposes of the present chapter, we define the polar as $C^\circ \triangleq \{z \mid \langle z, y \rangle \leq 0 \text{ for all } y \in C\}$. The earlier definition in Section 2.2 gives the same result for cones, justifying the notation.

Following Rockafellar and Wets [1998], recall that the (contingent) *tangent cone* to a set $C \subset \mathbb{R}^m$ at $x \in C$ is defined as

$$T_C(x) \triangleq \limsup_{\tau \searrow 0} (C - x)/\tau = \{\Delta x \mid x + \tau \Delta x' \in C, \tau \searrow 0, \Delta x' \to \Delta x\}.$$

This agrees with the tangent cone of convex analysis in that case, justifying the notation. The set of *regular normals* is defined as the polar $\widehat{N}_C(x) \triangleq T_C(x)^\circ$. The set of *normals in the general sense* is defined as $N_C(p, d) \triangleq \limsup \widehat{N}_C(p', d')$. Since we will be dealing with closed sets, it suffices to define the set of *regular tangents* $\widehat{T}_C(p, d)$ as the polar of this cone [Ibid., Theorem 6.28]. We always have $T_C(p, d) \supset \widehat{T}_C(p, d)$ [Ibid., Theorem 6.26]. The set C is *regular* at (p, d) when equality holds.

The following results will be of frequent use. As should be clear from the context, F^{-1} sometimes denotes the set-valued inverse, which always exists.

Theorem 3.1. [Ibid., Theorem 6.31] *For closed sets $X \subset \mathbb{R}^n$ and $D \subset \mathbb{R}^m$, a C^1 mapping $F : \mathbb{R}^n \to \mathbb{R}^m$, and the set $C \triangleq \{x \in X \mid F(x) \in D\}$, we have $T_C(x) \subset \{w \in T_X(x) \mid \nabla F(x) w \in T_D(F(x))\}$. Also $\widehat{T}_C(x) \supset \{w \in \widehat{T}_X(x) \mid \nabla F(x) w \in \widehat{T}_D(F(x))\}$, subject to the constraint qualification*

$$y \in N_D(F(x)),\ 0 \in \nabla F(x)^* y + N_X(x) \implies y = 0. \tag{3.3}$$

Theorem 3.2. [Ibid., Theorem 6.14] *Analogously to the above, $\widehat{N}_C(x) \supset \{\nabla F(x)^* y + z \mid y \in \widehat{N}_D(F(x)), z \in \widehat{N}_X(x)\}$, and subject to* (3.3)*, $N_C(x) \subset \{\nabla F(x)^* y + z \mid y \in N_D(F(x)), z \in N_X(x)\}$.*

Theorem 3.3. [Ibid., Theorem 6.43] *For $D \triangleq F(X)$, we have the inclusion $T_D(u) \supset \bigcup_{x \in F^{-1}(u) \cap X} \nabla F(x) T_X(x)$. Subject to there existing a neighbourhood $U \ni u$ such that $F^{-1}(U) \cap X$ is bounded, also $\widehat{T}_D(u) \supset \bigcap_{x \in F^{-1}(u) \cap X} \nabla F(x) \widehat{T}_X(x)$.*

Taking the tangent to the graph of a set-valued function S at $(y, z), z \in S(y)$, we get the (contingent) graphical derivative

$$\begin{aligned} DS(y|z)(\Delta y) &\triangleq \{\Delta z \mid (\Delta y, \Delta z) \in T_{\operatorname{Graph} S}(y, z)\} \\ &= \{\Delta z \mid z + \tau \Delta z' \in T(y + \tau \Delta y'), \tau \searrow 0, (\Delta y', \Delta z') \to (\Delta y, \Delta z)\}. \end{aligned}$$

Likewise we define the regular graphical derivative $\widehat{D}S$ from $\widehat{T}_{\operatorname{Graph} S}$. The mapping S is said to be graphically regular at (y, z) if its graph is regular at this point.

3.2.2 Euclidean Jordan algebras

In this subsection we introduce the bare minimum of the theory of (finite-dimensional Euclidean) Jordan algebras necessary for the analysis of this thesis. We will rely on the Jordan algebra of quadratic forms related to the familiar second-order cone as a concrete example in our exposition. More detailed treatment may be found in, e.g., Faraut and Korányi [1994] and Koecher [1999].

A (real) Jordan algebra $\mathcal{J}$ is a real vector space endowed with a multiplication operator $\circ : \mathcal{J} \times \mathcal{J} \to \mathcal{J}$, that is bilinear, commutative, and satisfies the property

$$x \circ (x^2 \circ y) = x^2 \circ (x \circ y) \text{ where } x^2 = x \circ x.$$

We assume in addition that $\mathcal{J}$ is *Euclidean* (or *formally real*), satisfying: $x^2 + y^2 = 0$ implies $x = y = 0$.

Then $\mathcal{J}$ has a multiplicative unit element e ($x \circ e = x$). An element x is called invertible, if there exists an element x^{-1}, such that $x \circ x^{-1} = x^{-1} \circ x = e$. We denote by $L(x)$ the symmetric linear operator $(x \circ \cdot) : \mathcal{J} \to \mathcal{J}$. The operator $L(x)$ is invertible precisely when x is. We say that x and y *operator-commute* when $L(x)L(y) = L(y)L(x)$.

An element c is called an *idempotent*, if $c \circ c = c$. It is *primitive*, if it cannot be composed by summing from other idempotents. A *complete orthogonal system of primitive idempotents* or a *Jordan frame* $c_1, \ldots, c_r$ is such that $c_i \circ c_j = 0$ for $i \neq j$, and $\sum_{j=1}^{r} c_j = e$. The number r is the *rank* of $\mathcal{J}$.

It turns out that for each $x \in \mathcal{J}$, there exist unique real numbers $\zeta_1, \ldots, \zeta_r$, called the *eigenvalues* of x, and a Jordan frame $c_1, \ldots, c_r$, such that $x = \sum_{j=1}^{r} \zeta_i c_i$. If all the eigenvalues are positive, x is called *positive-definite*. The number of non-zero eigenvalues is the *rank* of x. Powers of x may be defined as $x^\alpha \triangleq \sum_j \zeta_i^\alpha c_i$ when meaningful. We may also define the determinant $\det x \triangleq \prod_j \zeta_j$, and the trace $\operatorname{tr} x \triangleq \sum_j \zeta_j$.

The trace may be used to define the inner product $\langle x, y \rangle \triangleq \operatorname{tr}(x \circ y)$, which is positive-definite and associative, satisfying $\langle L(x)y, z \rangle = \langle y, L(x)z \rangle$. We may also define the norms $\|x\|_F \triangleq \sqrt{\sum_j \zeta_j^2} = \sqrt{\langle x, x \rangle}$ and $\|x\|_2 \triangleq \max_j |\zeta_j|$. According to [Schmieta and Alizadeh, 2001, Lemma 4], we have $\|x \circ y\|_F \leq \|x\|_2 \|y\|_F \leq \|x\|_F \|y\|_F$.

The *quadratic presentation* of x is defined as $Q_x \triangleq 2L(x)^2 - L(x^2)$. It turns out that the invertibility of x is equivalent to the invertibility of Q_x as well. Important properties, which can be found in Schmieta and Alizadeh [2003], include $Q_x^k = Q_{x^k}$, $Q_{Q_x y} = Q_x Q_y Q_x$, $Q_x x^{-1} = x$, and $Q_x e = x^2$.

Also denote $Q_{x,y} \triangleq L(x)L(y) + L(y)L(x) - L(x \circ y)$. Then $Q_x = Q_{x,x}$. For a Jordan frame $c_1, \ldots, c_r$, $Q_{c_i, c_j} = 2L(c_i)L(c_j) = 2L(c_j)L(c_i)$ for $i \neq j$, and the operators Q_{c_i} ($i = 1 \ldots r$) and $2Q_{c_i, c_j}$ ($i < j$) form a complete set of orthogonal projection operators in $\mathcal{J}$. More precisely, $\mathcal{R}(Q_{c_i}) = \{x \mid L(c_i)x = x\} = \mathbb{R}c_i$ and

$\mathcal{R}(Q_{c_i,c_j}) = \{x \mid L(c_i)x = L(c_j)x = x/2\}$ for $i \neq j$, as follows from the theory of *Peirce decompositions*. If $x = \sum_{i=1}^{r} \zeta_i c_i$, then $L(x) = \sum_i \zeta_i Q_{c_i} + \sum_{i<j}(\zeta_i + \zeta_j)Q_{c_i,c_j} = \sum_{i,j}(\zeta_i + \zeta_j)Q_{c_i,c_j}/2$.

Example 3.1 (Quadratic forms)**.** Consider the space $\mathcal{E}_{m+1}$ of $m+1$ element vectors $x = (x^0, \bar{x})$ with $x^0 \in \mathbb{R}$ and $\bar{x} \in \mathbb{R}^m$. Define the operator $\circ$ on $\mathcal{E}_{m+1}$ as

$$x \circ y = (x^T y, x^0 \bar{y} + y^0 \bar{x}).$$

Then $(\mathcal{E}_{m+1}, \circ)$ is a Euclidean Jordan algebra with inner product $\langle x, y \rangle = 2x^T y$, identity $e = (1, 0)$, and rank $r = 2$. The operator $L(x)$ is given by

$$L(x) = \operatorname{Arw}(x) \triangleq \begin{bmatrix} x^0 & \bar{x}^T \\ \bar{x} & x^0 I \end{bmatrix}$$

with I the identity matrix. Denote $R \triangleq \left[\begin{smallmatrix} 1 & 0 \\ 0 & -I \end{smallmatrix}\right]$. Then $\det x = x^T R x = (x^0)^2 - \|\bar{x}\|^2$, and $x^{-1} = Rx/\det x$ when $\det x \neq 0$.

3.2.3 Symmetric cones

The *cone of squares* of $\mathcal{J}$ is defined as $\mathcal{K} = \mathcal{K}(\mathcal{J}) \triangleq \{x^2 \mid x \in \mathcal{J}\}$. It turns out that the cones generated this way are precisely the so-called symmetric cones[1], and are the same as the self-scaled cones of Nesterov and Todd [1997]. Important properties include [Faraut and Korányi, 1994; Koecher, 1999]

(i) $\operatorname{int}\mathcal{K} = \{x \in \mathcal{J} \mid x \text{ is positive-definite}\} = \{x \in \mathcal{J} \mid L(x) \text{ pos. def.}\}$.

(ii) $\langle x, y \rangle \geq 0$ for all $y \in \mathcal{K}$ if and only if $x \in \mathcal{K}$, and

(iii) $\langle x, y \rangle > 0$ for all $y \in \mathcal{K} \setminus \{0\}$ if and only if $x \in \operatorname{int}\mathcal{K}$.

(iv) Q_x for $x \in \operatorname{int}\mathcal{K}$ maps $\mathcal{K}$ onto itself.

(v) For $x, y \in \operatorname{int}\mathcal{K}$, there is a unique $a \in \operatorname{int}\mathcal{K}$, such that $x = Q_a y$.

(vi) For any $x, y \in \mathcal{K}$, $\langle x, y \rangle = 0$ if and only if $x \circ y = 0$ [Faybusovich, 1997b].

In relation to (barrier) interior point methods, the following properties are particularly important:

(vii) $B(x) \triangleq -\log(\det x)$ tends to infinity as x goes to $\operatorname{bd}\mathcal{K}$.

(viii) $\nabla B(x) = -x^{-1}$, $\nabla^2 B(x) = Q_x$ when differentiated with respect to $\langle \cdot, \cdot \rangle$.

(ix) $\|y\|_x \triangleq \|Q_x^{-1/2} y\|_F$ defines a local norm around $x \in \operatorname{int}\mathcal{K}$, such that $\|y - x\|_x = \|Q_x^{-1/2} y - e\|_F \leq 1$ implies $y \in \mathcal{K}$. (This follows by considering the eigenvalue definition of $\|\cdot\|_F$, and the onto-property of Q_x; cf. also Nesterov and Todd [1997].)

[1] The term is also used of just $\operatorname{int}\mathcal{K}$.

Next we consider the normal structure of $\mathcal{K}$. Recall that, following Hiriart-Urruty and Lemaréchal [1993], the set of *ϵ-normals* with respect to the inner product $\langle \cdot, \cdot \rangle$ to a convex set $C \subset \mathbb{R}^m$ at x is defined as

$$N_{C,\epsilon}(x) \triangleq \{s \in \mathbb{R}^m \mid \langle s, y \rangle \leq \langle s, x \rangle + \epsilon \text{ for all } y \in C\}, \quad \epsilon \geq 0.$$

This definition reduces to the usual normal cone N_C at $\epsilon = 0$. In $\mathcal{J}$ we use the trace-based inner product. When $\mathcal{J}$ is simple, i.e., contains no non-trivial ideal, every associative symmetric bilinear form on $\mathcal{J}$ is given by a constant factor times this inner product, and therefore a different choice only scales the ϵ-normal set. Otherwise $\mathcal{J}$ is a product of simple Jordan algebras, and we get matrix scaling.

Lemma 3.1. *For $\epsilon \geq 0$, $N_{\mathcal{K},\epsilon}(x) \triangleq -\{s \in \mathcal{K} \mid \langle x, s \rangle \leq \epsilon\}$. This may be written $-Q_x^{-1/2}\{z \in \mathcal{K} \mid \operatorname{tr} z \leq \epsilon\}$ for $x \in \operatorname{int} \mathcal{K}$. Furthermore, for $\epsilon = 0$, s is a non-negatively weighted sum of those primitive idempotents in any Jordan frame of x with zero eigenvalue. Thus $p \circ N_{\mathcal{K}}(p) = 0$.*

Proof. For a cone, we must have $\langle s, y \rangle \leq 0$ in the definition of $N_{\mathcal{K},\epsilon}(x)$, for otherwise a scaling of y would violate the inequality $\langle s, y \rangle \leq \langle s, x \rangle + \epsilon$. But for a symmetric cone, $\langle s, y \rangle \leq 0$ for all $y \in \mathcal{K}$ implies $-s \in \mathcal{K}$. Choosing $y = 0$, we also get $\langle s, x \rangle + \epsilon \geq 0$. This says $-\langle x, s \rangle = -\operatorname{tr}(Q_x^{1/2} s) \leq \epsilon$. Since $Q_x^{1/2}$ maps $\mathcal{K}$ onto $\mathcal{K}$ for $x \in \operatorname{int} \mathcal{K}$, $N_{\mathcal{K},\epsilon}$ can be expressed as claimed by negating s.

Let then $\epsilon = 0$, $\langle s, x \rangle = 0$, and $x = \sum_{j=1}^r \zeta_j c_j$ for the primitive idempotents c_j and $\zeta_j \geq 0$. Since all these elements are in $\mathcal{K}$, $\zeta_j \langle s, c_j \rangle = 0$ for all j. This says that $\langle s, c_i \rangle = 0$, and therefore $s \circ c_i = 0$, for all c_i with $\zeta_i > 0$. The same must hold for the primitive idempotents in the decomposition of s by the properties (ii) and (iii) listed above. □

Example 3.2 (The second order cone)**.** For the Jordan algebra $\mathcal{E}_{m+1}$ of quadratic forms, considered in Example 3.1, we get the so-called second order cone, $\mathcal{K} = \{x \mid x^0 \geq \|\bar{x}\|\}$. For $0 \neq x \in \operatorname{bd} \mathcal{K}$, $x^2 = 2x^0 x$, so scaling gives a primitive idempotent, and the only orthogonal one is proportional to Rx. Therefore, $N_{\mathcal{K}}(x) = \{-\alpha Rx \mid \alpha \geq 0\}$, a set approximated by $\{-\alpha y^{-1} \mid \alpha \geq 0\}$ for invertible y close to x.

3.3 ϵ-complementary pairs in a symmetric cone

We next consider the tangent and normal structure of the graph of $N_{\mathcal{K}_\epsilon}$.

Definition 3.1. We say that two elements $p, d \in \mathcal{K}$ are *strictly complementary*, if $p \circ d = 0$, and $p + d \in \operatorname{int} \mathcal{K}$ [Pataki, 1996; Schmieta and Alizadeh, 2003].

Lemma 3.2.

(i) Suppose that p, d are strictly complementary. Then $p \circ \Delta d + d \circ \Delta p = 0$ if and only if $(\Delta p, \Delta d) = (L(p)\eta, -L(d)\eta)$ for some $\eta \in \mathcal{J}$.

(ii) When the latter representation of (i) holds, $\langle \Delta p, N_{\mathcal{K}}(p)\rangle = \langle \Delta d, N_{\mathcal{K}}(d)\rangle = 0$*, and consequently* $\Delta p \in T_{\mathcal{K}}(p)$, $\Delta d \in T_{\mathcal{K}}(d)$.

(iii) The consequences of (ii) respectively imply that $\operatorname{tr}(p \circ \Delta d + d \circ \Delta p) = 0$*, and* $\operatorname{tr}(p \circ \Delta d + d \circ \Delta p) \geq 0$.

Proof. (i) Since $p \circ d = 0$, there exists a common Jordan frame $c_1, \ldots, c_r$ and eigenvalues $\zeta_1, \ldots, \zeta_r \geq 0$ and $\sigma_1, \ldots, \sigma_r \geq 0$ with $\zeta_i \sigma_i = 0$ and $\zeta_i + \sigma_i > 0$, such that $p = \sum_i \zeta_i c_i$, and $d = \sum_i \sigma_i c_i$. Therefore, recalling the representation of $L(p) = \sum_{i,j} (\zeta_i + \zeta_j) Q_{c_i,c_j}/2$ and $L(d) = \sum_{i,j} (\sigma_i + \sigma_j) Q_{c_i,c_j}/2$, we have

$$L(p)\Delta d + L(d)\Delta p = 0 \iff Q_{c_i,c_j}\big((\zeta_i + \zeta_j)\Delta d + (\sigma_i + \sigma_j)\Delta p\big) = 0 \text{ for all } i, j.$$

Note that always either $\zeta_i + \zeta_j > 0$ or $\sigma_i + \sigma_j > 0$, so that $\zeta_i + \zeta_j = 0$ forces $Q_{c_i,c_j}\Delta p = 0$, and the other way around. Consequently, Δp is proportional to Δd on $\mathcal{R}(Q_{c_i,c_j})$. Therefore $\Delta p, \Delta d \propto Q_{c_i,c_j}\eta$ for some $\eta \in \mathcal{J}$, which may be chosen the same for all i, j by orthogonality of the projection operators Q_{c_i,c_j}. The correct proportionality factors are given by the choice $\Delta p = L(p)\eta$ and $\Delta d = -L(d)\eta$ for some $\eta \in \mathcal{J}$.

On the other hand, strictly complementary p and d operator-commute (as seen from the Q-decomposition of L; cf. [Schmieta and Alizadeh, 2003, Theorem 27]), so the equality follows from the representation.

(ii) As a consequence of the representation of (i), $\langle N_{\mathcal{K}}(p), \Delta p\rangle = \langle p \circ N_{\mathcal{K}}(p), \eta\rangle = 0$, from which also $\Delta p \in T_{\mathcal{K}}(p)$. The claims for Δd follow similarly.

(iii) From Lemma 3.1, $-p \in N_{\mathcal{K}}(d)$ and $-d \in N_{\mathcal{K}}(p)$. The claim follows. □

Theorem 3.4. *Let* $C_\epsilon \triangleq \{(p, d) \in \mathcal{K} \times \mathcal{K} \mid \operatorname{tr} p \circ d \leq \epsilon\}$ *and* $(p, d) \in C_\epsilon$*. Then, for* $\epsilon > 0$, C_ϵ *is regular at* (p, d)*, and*

$$T_{C_\epsilon}(p, d) = \{(\Delta p, \Delta d) \in T_{\mathcal{K}}(p) \times T_{\mathcal{K}}(d) \mid \operatorname{tr}(p \circ \Delta d + d \circ \Delta p) \leq \infty(\epsilon - \operatorname{tr} p \circ d)\},$$

whereas for $\epsilon = 0$,

$$T_{C_0}(p, d) \subset \{(\Delta p, \Delta d) \in T_{\mathcal{K}}(p) \times T_{\mathcal{K}}(d) \mid p \circ \Delta d + d \circ \Delta p = 0\}, \text{ and}$$
$$\widehat{T}_{C_0}(p, d) \supset \{(L(p)\eta, -L(d)\eta) \mid \eta \in \mathcal{J}\}.$$

When p *and* d *are strictly complementary, the two right hand sides above are equal, and we have regularity as well as the representation*

$$T_{C_0}(p, d) = \{(\Delta p, \Delta d) \in \mathcal{J} \times \mathcal{J} \mid p \circ \Delta d + d \circ \Delta p = 0\}.$$

Proof. Write $C_\epsilon = \{(p, d) \in X \mid F(p, d) \in D\}$ for $X = \mathcal{K} \times \mathcal{K}$, $F(p, d) \triangleq p \circ d$, and $D \triangleq \{x \in \mathcal{J} \mid \operatorname{tr} x \leq \epsilon\}$. Then $N_D(x) = e\{\alpha \geq 0 \mid \alpha(\operatorname{tr} x - \epsilon) = 0\}$ and $\nabla F(p, d) = (L(d), L(p))$. Suppose $\epsilon > 0$. That T_{C_ϵ} is included in the set of the statement is now immediate from Theorem 3.1. Since the convex sets D and $\mathcal{K}$

are regular, the same theorem provides the other inclusion and hence equality and regularity after we check the constraint qualification (3.3), i.e.,

$$\left.\begin{array}{l}(z_p, z_d) \in N_{\mathcal{K}}(p) \times N_{\mathcal{K}}(d),\ \alpha e \in N_D(F(p,d)) \\ (z_p, z_d) + \alpha \nabla F(p,d)^* e = 0\end{array}\right\} \implies \alpha = 0.$$

The sum condition writes out to $z_p + \alpha d = 0$ and $z_d + \alpha p = 0$. Multiplying from left by p and d, respectively, we get applying Lemma 3.1 that $\alpha(p \circ d) = 0$ in both cases. Since $\epsilon > 0$, either $\alpha = 0$ or $\operatorname{tr} p \circ d = \epsilon$. But in the latter case $p \circ d \neq 0$, which provides a contradiction.

Let now $\epsilon = 0$. For the first inclusion, note that we may take $D = \{0\}$, and then apply again Theorem 3.1. As for the second inclusion, the constraint qualification is not satisfied this time, so some extra tricks that furbish a similar aid are needed (only yielding a lower bound in the non-strictly complementary case). So let $U(p,d) \triangleq \{(L(p)\eta, -L(d)\eta) \mid \eta \in \mathcal{J}\}$. Now, for the polar of this cone we have

$$\begin{aligned} U^\circ(p,d) &= \{(z_p, z_d) \mid \langle (z_p, z_d), (\Delta p, \Delta d)\rangle \leq 0, \text{ for all } (\Delta p, \Delta d) \in U(p,d)\} \\ &= \{(z_p, z_d) \mid \langle (z_p, z_d), (L(p)\eta, -L(d)\eta)\rangle \leq 0 \text{ for all } \eta\} \\ &= \{(z_p, z_d) \mid \langle L(p)z_p - L(d)z_d, \eta\rangle \leq 0 \text{ for all } \eta\} \\ &= \{(z_p, z_d) \mid L(p)z_p - L(d)z_d = 0\}. \end{aligned}$$

Let then $v \in \widehat{N}_{C_0}(p,d)$. We will show that $v \in U^\circ(p,d)$. Following the argument of the proof of Theorem 3.2 in [Rockafellar and Wets, 1998, Theorem 6.14], we choose a smooth function h with $\arg\max_{C_0} h = \{(p,d)\}$, $\nabla h(p,d) = v$, existent by [Ibid., Theorem 6.11]. Then we consider the penalty functions

$$\psi_{[k]}(p,d,u) \triangleq -h(p,d) + \frac{1}{2\tau_{[k]}}\|F(p,d) - u\|^2, \quad \tau_{[k]} \searrow 0, \quad k = 1, 2, \ldots.$$

Minimising these functions on $\mathcal{K} \times \mathcal{K} \times D$ yields convergent sequences

$$h_{[k]} = z_{[k]} + \nabla F(p_{[k]}, d_{[k]})^* y_{[k]} \to v, \quad (p_{[k]}, d_{[k]}) \in \mathcal{K} \times \mathcal{K} \to (p,d),$$

with $h_{[k]} \triangleq \nabla h(p_{[k]}, d_{[k]})$, $z_{[k]} = (z_{p,[k]}, z_{d,[k]}) \in N_{\mathcal{K}}(p_{[k]}) \times N_{\mathcal{K}}(d_{[k]})$, $y_{[k]} = \alpha_{[k]} e \in N_D(u_{[k]})$, and $u_{[k]} \in D$.

Normally the constraint qualification is used to prove that v is of the required form, but in this case we need another argument. Presently

$$h_{[k]} = (z_{p,[k]} + \alpha_{[k]} d_{[k]}, z_{d,[k]} + \alpha_{[k]} p_{[k]}). \tag{3.4}$$

Multiplying (3.4) by $(L(p_{[k]}); -L(d_{[k]}))$ yields $\alpha_{[k]}(L(p_{[k]})d_{[k]} - L(d_{[k]})p_{[k]}) = 0$. On the other hand, since $h_{[k]} \to v$ is bounded, and $p_{[k]} - p \to 0$, we have $\|(L(p) - L(p_{[k]}))(z_{p,[k]} + \alpha_{[k]} d_{[k]})\| \to 0$, and likewise for the other term. Therefore

$$\begin{aligned} (L(p); -L(d))v &= \lim_k (L(p); -L(d))h_{[k]} \\ &= \lim_k (L(p) - L(p_{[k]}); -L(d) + L(d_{[k]}))h_{[k]} = 0, \end{aligned} \tag{3.5}$$

and consequently $U^\circ(p,d) \supset \widehat{N}_{C_0}(p,d)$.

An argument similar to (3.5) shows that $U^\circ(p,d)$ is outer semicontinuous. Therefore $U^\circ(p,d) \supset N_{C_0}(p,d)$, and consequently $U^{\circ\circ}(p,d) \subset \widehat{T}_{C_0}(p,d)$. As clearly $U(p,d) \subset U^{\circ\circ}(p,d)$, we get that part of our claim.

In the strictly complementary case, we just apply Lemma 3.2 to the presentations obtained for the general case. □

Corollary 3.1. *For $\epsilon > 0$,*

$$\widehat{N}_{C_\epsilon}(p,d) = \{(N_{\mathcal{K}}(p), N_{\mathcal{K}}(d)) + \alpha(d,p) \mid \alpha \geq 0, \alpha(\epsilon - \operatorname{tr} p \circ d) = 0\}.$$

For $\epsilon = 0$,

$$\begin{aligned} &\widehat{N}_{C_0}(p,d) \supset \{(N_{\mathcal{K}}(p) + L(d)\eta, N_{\mathcal{K}}(d) + L(p)\eta) \mid \eta \in \mathcal{J}\}, \\ &N_{C_0}(p,d) \subset \{(z_p, z_d) \mid L(p)z_p = L(d)z_d\}. \end{aligned} \tag{3.6}$$

When p and d are strictly complementary, the two right hand sides above are equal, and we also have $\widehat{N}_{C_0}(p,d) = \{(L(d)\eta, L(p)\eta) \mid \eta \in \mathcal{J}\}$.

Proof. For $\epsilon > 0$ and the first inclusion for $\epsilon = 0$, the claims follow by applying Theorem 3.2 instead of Theorem 3.1 in the proof of Theorem 3.4, or alternatively through polarity relationships. The latter inclusion for $\epsilon = 0$ follows directly from the proof of Theorem 3.4. The claim on the strictly complementary case follows from applying Lemma 3.2(i) to the expression in (3.6). □

3.4 The class of functions

3.4.1 A class of convex functions

We now consider convex functions on $\mathbb{R}^m$ of the form (3.1). That is,

$$f(y) \triangleq \sup\{\langle B^*y + c, p\rangle \mid p \in \mathcal{K}, Ap = b\} = \sigma(B^*y + c; V), \tag{3.7}$$

where $\mathcal{K}$ is a symmetric cone with associated Jordan algebra $\mathcal{J}$, $A : \mathcal{J} \to \mathbb{R}^{m_A}$ and $B : \mathcal{J} \to \mathbb{R}^m$ are linear mappings, $c \in \mathcal{J}$, $V \triangleq \{p \in \mathcal{K} \mid Ap = b\}$, and $\sigma(\cdot; V)$ is the support function of V. We require that $\mathcal{N}(B^*; A^*) = \{0\}$, and that

$$A\mathcal{K} = \{\lambda \in \mathbb{R}^{m_A} \mid \lambda \geq 0\}, \tag{3.8}$$

$$\mathcal{N}(A) \cap \mathcal{K} = \{0\}, \text{ and} \tag{3.9}$$

$$b \in A(\operatorname{int} \mathcal{K}). \tag{3.10}$$

Example 3.3 (Euclidean norms)**.**

(i) If $\mathcal{K}$ is the second-order cone on $\mathcal{E}_{m+1}$, $Ap \triangleq p^0 = \langle e/2, p\rangle$ (recalling that the inner product on $\mathcal{E}_{m+1}$ is two times the standard $\mathbb{R}^{m+1}$ inner product) $b \triangleq 1$, $c \triangleq (0, -a/2)$, and $Bp \triangleq \bar{p}$ (whence $B^*y = (0, y/2)$), we get $f(y) = \sup\{(y-a)^T\bar{p} \mid 1 = p^0 \geq \|\bar{p}\|\} = \|y - a\|$.

(ii) Weighted sums $\sum_{k=1}^{n} \|W_k(y - a_k)\|$ of Euclidean norms can be represented by a straightforward extension: $p = (p_1, \ldots, p_n) \in \mathcal{K}^n$, $Ap \triangleq (p_1^0, \ldots, p_n^0)$, $b \equiv 1$, $B^*y \triangleq ((0, W_1 y), \ldots, (0, W_n y))/2$, and $c \triangleq -((0, W_1 a_1), \ldots, (0, W_n a_n))/2$.

(iii) Finally, if we instead set $Ap \triangleq \sum_{k=1}^{n} p_k^0$ and $b = 1$, the supremum favours maximum $\langle W_k(y - a_k), \bar{p} \rangle$. We therefore have $f(y) = \max_{k=1,\ldots,n} \|W_k(y - a_k)\|$.

Lemma 3.3. *We have the following equivalences and implications:*

(i) Assumption (3.8) *is equivalent to $A^*\lambda \in \mathcal{K}$ if and only if $\lambda \geq 0$.*

(ii) Under assumptions (3.8) *and* (3.10)*, we have $b > 0$, and assumption* (3.9) *is equivalent to V being non-empty and bounded.*

(iii) Assumption (3.8) *is equivalent to $(Ap)_i = \langle a_i, p \rangle$ for some $a_i \in \mathcal{K} \setminus \{0\}$, $\langle a_i, a_j \rangle = 0$ $(j \neq i)$.*

(iv) Under assumption (3.8)*, assumption* (3.9) *is equivalent to $\sum_{i=1}^{m_A} a_i \in \operatorname{int} \mathcal{K}$.*

Proof. (iii) Since $p \mapsto (Ap)_i$ is linear, there exists a vector a_i such that $(Ap)_i = \langle a_i, p \rangle$. It then follows from assumption (3.8) that $\langle a_i, p \rangle \geq 0$ for all $p \in \operatorname{int} \mathcal{K}$ and $i = 1, \ldots, m_A$. By surjectivity $a_i \neq 0$, so that $a_i \in \mathcal{K} \setminus \{0\}$. If a_i and a_j were not orthogonal, we could write $a_j = \alpha a_i + w$ for $\alpha > 0$ and w orthogonal to a_i. If $w \in \mathcal{K}$, $\langle a_i, p \rangle = \lambda_i$ for $p \in \mathcal{K}$ forces $\langle a_j, p \rangle \geq \alpha \lambda_i$, so surjectivity fails. If $w \notin \mathcal{K}$, the set $\mathcal{K} \cap (a_i + \mathbb{R}_+ w)$ is bounded, so again surjectivity fails.

(iv) Assumption (3.9) then says that there is no $v \in \mathcal{K}$ such that $\langle a_i, v \rangle = 0$ for all $i = 1, \ldots, m_A$. Since $\langle a_i, v \rangle \geq 0$, this is equivalent to $\langle \sum_i a_i, v \rangle > 0$ for all $v \in \mathcal{K}$, which says that $\sum_i a_i \in \operatorname{int} \mathcal{K}$.

(i) Let $p \in \mathcal{K}$. Then $Ap \geq 0$ is equivalent to $\langle Ap, \lambda \rangle \geq 0$ for all $\lambda \geq 0$, which says the same as $\langle p, A^*\lambda \rangle \geq 0$ for all $\lambda \geq 0$ and $p \in \mathcal{K}$. This is equivalent to $A^*\lambda \in \mathcal{K}$ for $\lambda \geq 0$. This shows the equivalence of $A\mathcal{K} \subset \{\lambda \geq 0\}$ with $A^*\{\lambda \geq 0\} \subset \mathcal{K}$.

Suppose then that $\lambda \not\geq 0$. Then for some $\lambda' \geq 0$, we have $\langle \lambda, \lambda' \rangle < 0$. When assumption (3.8) holds, $\lambda' = Ap'$ for some $p' \in \mathcal{K}$. Consequently $\langle p', A^*\lambda \rangle = \langle Ap', \lambda \rangle < 0$, so that $A^*\lambda \notin \mathcal{K}$.

As for the converse, if the inclusion $A\mathcal{K} \subset \{\lambda \geq 0\}$ is strict, then because these sets are closed convex cones, there exists some $\lambda' \in (-N_{A\mathcal{K}}(0)) \setminus \{\lambda \geq 0\}$. This says that for all $p \in \mathcal{K}$, $0 \leq \langle Ap, \lambda' \rangle = \langle p, A^*\lambda' \rangle$. Therefore $A^*\lambda' \in \mathcal{K}$ for $\lambda' \not\geq 0$, and consequently the condition $A^*\lambda \in \mathcal{K}$ if and only if $\lambda \geq 0$ cannot hold.

(ii) Assumption (3.10) and the representation of (iii) give $b > 0$. By the same assumption V is non-empty.

Suppose then that assumption (3.9) holds and $p_{[0]} \in V$. If V is unbounded, there is a sequence $z_{[k]} \in \mathcal{N}(A)$ $(k = 1, 2, \ldots)$, such that $p_{[k]} \triangleq p_{[0]} + z_{[k]} \in \mathcal{K}$, and $\|z_{[k]}\| \to \infty$. Now, we have $z \triangleq \lim p_{[k]}/\|p_{[k]}\| = \lim z_{[k]}/\|z_{[k]}\| \in \mathcal{K} \cap \mathcal{N}(A)$, as well as $\|z\| = 1$, which is a contradiction. Thus V is bounded.

On the other hand, if $0 \neq z \in \mathcal{N}(A) \cap \mathcal{K}$, and $p \in V$, then $p + \lambda z \in V$ for all $\lambda \geq 0$, so that V is unbounded. □

Our next task is to calculate $\partial_\epsilon f(y)$. Towards that end, we first need to study $N_{V,\epsilon}$. As is well known, actually $N_{V,\epsilon}(p) = \partial_\epsilon \delta_V(p)$ for the indicator function of the set V. In the present case, $\delta_V = \delta_{\mathcal{K}} + \delta_{\{p \in \mathcal{J} | Ap = b\}}$. Therefore, as the relative interior of V is non-empty by assumption (3.10), we may apply [Hiriart-Urruty and Lemaréchal, 1993, Theorem XI.3.1.1] to yield $N_{V,\epsilon}(p) = N_{\mathcal{K},\epsilon}(p) + \mathcal{R}(A^*)$, the ϵ-normal set of a linear space being the normal set. Thus by Lemma 3.1, $z \in N_{V,\epsilon}(p)$ at $p \in V$ iff for some $\lambda \in \mathbb{R}^{m_A}$ and $d \in \mathcal{K}$, we have $\langle p, d \rangle \leq \epsilon$ and $z + A^*\lambda + d = 0$.

Now, applying [Ibid., Theorem XI.3.2.1 and Example XI.1.2.5] in the first two equalities, we get

$$\begin{aligned}
\partial_\epsilon f(y) &= B \partial_\epsilon \sigma(B^* y + c; V) \\
&= B\{p \in V \mid \sigma(B^* y + c; V) \leq \langle p, B^* y + c \rangle + \epsilon\} \\
&= B\{p \in V \mid \langle p' - p, B^* y + c \rangle \leq \epsilon \text{ for all } p' \in V\} \\
&= B\{p \in V \mid B^* y + c \in N_{V,\epsilon}(p)\} \\
&= B\{p \in V \mid -d \in N_{\mathcal{K},\epsilon}(p), B^* y + A^*\lambda + d + c = 0\} \\
&= \{Bp \mid \langle p, d \rangle \leq \epsilon, Ap = b, B^* y + A^*\lambda + d + c = 0,\ p, d \in \mathcal{K}\}.
\end{aligned} \tag{3.11}$$

Remark 3.1. The set of equations for $0 \in \partial_\epsilon f(y)$ are very similar to the standard primal-dual equations for barrier methods, but without an explicit central path ($p \circ d = \mu e$) selected. Indeed, let $f^\mu(y) \triangleq \sup_{p \in V}\{\langle B^* y + c, p \rangle + \mu \log(\det p)\}$ be a barrier-smoothing of f. It is differentiable because $\log(\det p)$ is strictly concave in $\operatorname{int} \mathcal{K}$ (with $\nabla^2 \log(\det p) = -Q_p$), and we have $\nabla f^\mu(y) = B\{p \in V \mid B^* y + c + \mu p^{-1} \in -N_V(p)\} = \{Bp \mid Ap = b, B^* y + A^*\lambda + c + d = 0, p \circ d = \mu e, p, d \in \mathcal{K}\}$, using $d = \mu p^{-1}$.

After we look at the difference of functions of the form (3.7) shortly, we will be doing some second-order analysis, where we need the following notion of non-degeneracy. Conditions ensuring this will be further discussed in Section 3.4.5.

Definition 3.2. We say that a strictly complementary pair (p, d) is *non-degenerate relative to a subspace* $X \subset \mathcal{J}$, if $(L(d)\eta, L(p)\eta) \in \mathcal{R}(A^*) \times (X \cap \mathcal{N}(A))$ implies $\eta = 0$.

Example 3.4 (Euclidean norms)**.** Consider the base case of Example 3.3. At $y = a$, we have $d = 0$ and strict complementarity holds for $p = (1, \bar{p})$ with $\|\bar{p}\| < 1$. As $L(p)$ is non-singular, (p, d) is not non-degenerate (relative to $\mathcal{J}$), but it is non-degenerate relative to $\mathcal{N}(B) = \mathbb{R}e = \mathcal{R}(A^*)$.

3.4.2 Taking the difference

Let f be of the form (3.7), and subscript the data and variables as B_f, A_f, c_f, b_f, $\mathcal{K}_f$, etc. Let ν be another function in this class, with similar subscripts. Now let $f_\nu \triangleq f - \nu$, making f_ν a diff-convex function.

Example 3.5 (Location problems)**.** Recalling from Example 3.3 that sums and maxima of (matrix-scaled) Euclidean distances can be represented in the form (3.7), we find that, e.g., the multisource Weber problem objective function (1.5) has the form f_ν. So does the "MO" clustering objective and the reformulations of the Euclidean TSP, as discussed in Section 1.4 and to be studied in Chapters 6 and 7.

Our objective in Chapter 4 that follows is to minimise f_ν, or at least find an approximately critical point. That is, we are interested in finding ϵ-semi-critical points. This property we recall to be defined as

$$0 \in \partial_\epsilon^{\mathrm{DC}} f_\nu(y) \triangleq \bigcup\{\partial_{\epsilon_1} f(y) - \partial_{\epsilon_2} \nu(y) \mid \epsilon_1 + \epsilon_2 = \epsilon, \epsilon_1, \epsilon_2 \geq 0\}.$$

Now, note that the condition

$$\operatorname{tr} p_f \circ d_f \leq \epsilon_1 \text{ and } \operatorname{tr} p_\nu \circ d_\nu \leq \epsilon_2 \text{ for some } \epsilon_1 + \epsilon_2 = \epsilon, \epsilon_1, \epsilon_2 \geq 0$$

reduces to $\operatorname{tr}(p_f, p_\nu) \circ (d_f, d_\nu) \leq \epsilon$ in the product cone $\mathcal{K} \triangleq \mathcal{K}_f \times \mathcal{K}_\nu$. Thus, recalling the representation of $\partial_\epsilon f$ from (3.11), we actually get

$$\partial_\epsilon^{\mathrm{DC}} f_\nu(y) = \{B_- p \mid (p, d) \in C_\epsilon, Ap = b, B^* y + A^* \lambda + d + c = 0\},$$

with $A \triangleq (A_f, A_\nu)$, $B \triangleq (B_f; B_\nu)$, $B_- \triangleq (B_f; -B_\nu)$, $c \triangleq (c_f, c_\nu)$, and $b \triangleq (b_f, b_\nu)$.

Note that the non-degeneracy condition relative to $\mathcal{N}(B)$ is equivalent to that relative to $\mathcal{N}(B_-)$: supposing it did not hold for one, replacing $\eta = (\eta_f, \eta_\nu)$ with $(\eta_f, -\eta_\nu)$ in the definition, shows that it does not hold for the other, for $L(p)\eta \in \mathcal{N}(A)$ and $L(d)\eta \in \mathcal{R}(A^*)$ are unaffected by such change.

3.4.3 Second order behaviour

Lemma 3.4. *Let* $S_\epsilon \triangleq \{(p, d) \in C_\epsilon \mid Ap = b, B^* y + A^* \lambda + c + d = 0\}$. *Then*

$$T_{S_\epsilon}(p, d) \subset \left\{(\Delta p, \Delta d) \in T_{C_\epsilon}(p, d) \;\middle|\; A\Delta p = 0, B^* \Delta y + A^* \Delta\lambda + \Delta d = 0\right\},$$

with regularity and equality when $\epsilon > 0$, *or* p *and* d *are strictly complementary and non-degenerate relative to* $\mathcal{N}(B)$.

Proof. By (3.10), $b = Ap_0$ for some $p_0 \in \operatorname{int} \mathcal{K}$. So let $F(p, d) \triangleq (p, d) - (p_0, -c)$ and $D \triangleq \mathcal{N}(A) \times (\mathcal{R}(A^*) \cup \mathcal{R}(B^*))$. Then $S_\epsilon = \{(p, d) \in C_\epsilon \mid F(p, d) \in D\}$, and the inclusion for T_{S_ϵ} is again immediate from Theorem 3.1. For the equality we get the claim from the same theorem by proving the constraint qualification (3.3).

In the case of $\epsilon > 0$, applying Corollary 3.1, this constraint qualification becomes

$$z_p + \alpha d = A^* \lambda,\ z_d + \alpha p = s \in \mathcal{N}(A) \cap \mathcal{N}(B) \implies s = 0,\ \lambda = 0, \tag{3.12}$$

with $z_p \in N_{\mathcal{K}}(p)$, $z_d \in N_{\mathcal{K}}(d)$, $\alpha \geq 0$, and $\alpha(\epsilon - \langle p, d\rangle) = 0$.

Suppose $\alpha = 0$. Then, because $z_d \in -\mathcal{K}$, $s \in \mathcal{N}(A) \cap (-\mathcal{K}) = \{0\}$. Likewise, $A^*\lambda = z_p \in -\mathcal{K}$, whence by Lemma 3.3, $\lambda \leq 0$. Therefore, unless $\lambda = 0$, $0 = \alpha\langle p, z_p\rangle = \langle p, A^*\lambda\rangle = \langle b, \lambda\rangle < 0$, which is a contradiction.

Suppose then that $\alpha > 0$, whence $\langle p, d\rangle = \epsilon$. Recalling that $\langle p, z_p\rangle = \langle d, z_d\rangle = 0$ by Lemma 3.1, we get by multiplying the terms on the left hand side of (3.12) that $0 = \langle s, A^T\lambda\rangle = \langle z_p, z_d\rangle + \alpha^2\langle p, d\rangle = \langle z_p, z_d\rangle + \alpha^2\epsilon$. But this says that $\langle z_d, z_p\rangle < 0$, which is not possible, since $z_d, z_p \in -\mathcal{K}$.

When p and d are strictly complementary, Corollary 3.1 gives to (3.3) the format

$$L(d)\eta = A^*\lambda,\ L(p)\eta = s \in \mathcal{N}(A) \cap \mathcal{N}(B) \implies \lambda = 0,\ s = 0.$$

This is the non-degeneracy condition relative to $\mathcal{N}(B)$, because $L(p+d)$ is invertible by strict complementarity. □

Theorem 3.5. *Let G be a C^1 mapping with domain S_∞, such that it has a continuous partial inverse into $\{0\} \times \mathcal{R}(B^*)$.*[2] *Denote $G_\epsilon^{-1}(v) \triangleq \{(p, d) \in S_\epsilon \mid G(p, d) = v\}$. Then*

$$T_{GS_\epsilon}(v) \supset U_\epsilon^G(v) \triangleq \{\Delta v \in \nabla G(p, d)T_{S_\epsilon}(p, d) \mid (p, d) \in G_\epsilon^{-1}(v)\}.$$

Equality holds when for all $(p, d) \in G_\epsilon^{-1}(v)$,

$$(\Delta p, \Delta d) \in T_{S_\epsilon}(p, d),\ \nabla G(p, d)(\Delta p, \Delta d) = 0 \implies (\Delta p, \Delta d) = 0. \tag{3.13}$$

Proof. The inclusion $T_{GS_\epsilon}(v) \supset U_\epsilon^G(v)$ is just Theorem 3.3.

To show the equality, let $\Delta v \in T_{GS_\epsilon}(v)$. Then by definition there exists sequences $\Delta v_{[k]} \to \Delta v$ and $\tau_{[k]} \searrow 0$, as well as $(p_{[k]}, d_{[k]}) \in G_\epsilon^{-1}(v + \tau_{[k]}\Delta v_{[k]})$ $(k = 1, 2, \ldots)$. By Lemma 3.3(ii), $p_{[k]}$ is bounded. By the convergence of $\Delta v_{[k]}$ and the continuity of the partial inverse of G, $d_{[k]}$ is bounded in $\mathcal{R}(B^*)$. Since the remaining free part of $d_{[k]}$ is in $\mathcal{R}(A^*)$, it is bounded also in $\mathcal{N}(A)$. If $d_{[k]}$ were unbounded in $\mathcal{R}(A^*)$, we'd have $d_{[k]}\gamma_{[k]} \to A^*\lambda \in \mathcal{K} \setminus \{0\}$ for some $\gamma_{[k]} \searrow 0$. Consequently, since $\langle p_{[k]}, d_{[k]}\rangle \leq \epsilon$, we'd have $0 = \langle p_{[k]}, A^*\lambda\rangle = \langle b, \lambda\rangle$, in contradiction to the results of Lemma 3.3. Thus $(p_{[k]}, d_{[k]})$ is bounded, and by possibly moving to a subsequence, we may assume that $(p_{[k]}, d_{[k]}) \to (p, d)$ for some $(p, d) \in G_\epsilon^{-1}(v)$. We must also have $(p_{[k]} - p, d_{[k]} - d)/\gamma_{[k]} \to (\Delta p, \Delta d) \in T_{S_\epsilon}(p, d)$ for some sequence $\gamma_{[k]} \searrow 0$, with either $\gamma_{[k]} = \tau_{[k]}$, or $\tau_{[k]}/\gamma_{[k]} \to 0$ and $(\Delta p, \Delta d) \neq 0$. Thus also $(\tau_{[k]}/\gamma_{[k]})\Delta v_{[k]} = (G(p_{[k]}, d_{[k]}) - G(p, d))/\gamma_{[k]} \to \nabla G(p, d)(\Delta p, \Delta d)$. Therefore, if it can be chosen $\gamma_{[k]} = \tau_{[k]}$, we have the wanted representation for Δv. Otherwise

$$0 \neq (\Delta p, \Delta d) \in T_{S_\epsilon}(p, d),\ \nabla G(p, d)(\Delta p, \Delta d) = 0. \tag{3.14}$$

This is forbidden by the constraint qualification (3.13).[3] □

---

[2] In other words, the composition of projection into that subspace with any single-valued selection of the set-valued inverse of G is continuous.

[3] This constraint is related to that expressed in [Rockafellar and Wets, 1998, Theorem 4.26]. Indeed, this short direct proof is mostly about calculating the horizon limit supremum of $\bigcup_{(p,d)\in G_\epsilon^{-1}(v)}(S_\epsilon - (p, d))/\tau$.

Corollary 3.2. *Let* $G(p,d) \triangleq ((I;0)(B^*;A^*)^\dagger(-d-c), B_-p)$. *Then* $\partial_\epsilon^{\mathrm{DC}} f_\nu(y) = \{z \mid (y,z) \in GS_\epsilon\}$ *and* $D(\partial_\epsilon^{\mathrm{DC}} f_\nu)(y|z)(\Delta y) \supset U_\epsilon(y|z)$ *with*

$$U_\epsilon(y|z) \triangleq \bigcup_{(p,d)\in G_\epsilon^{-1}(y,z)} \{\Delta z \mid (\Delta y, \Delta z) \in \nabla G(p,d) T_{S_\epsilon}(p,d)\}.$$

Equality holds for $\epsilon = 0$ *when all* $(p,d) \in G_\epsilon^{-1}(y,z)$ *are strictly complementary and non-degenerate relative to* $\mathcal{N}(B_-)$.

Proof. The expression for $\partial_\epsilon^{\mathrm{DC}} f_\nu$ is just what we have shown in (3.11), written in a different form, as $G(p,d) = (y, B_-p)$ when $A^*\lambda + B^*y + d + c = 0$. This holds because we have from $\mathcal{N}(B^*;A^*) = \{0\}$ that $(B^*;A^*)^\dagger(B^*;A^*) = I$ and then

$$(I;0)(B^*;A^*)^\dagger(-d-c) = (I;0)(B^*;A^*)^\dagger(B^*;A^*)(y,\lambda) = y. \tag{3.15}$$

Since y continuously and uniquely determines d in $\mathcal{R}(B^*)$ through G, the partial inverse continuity condition of Theorem 3.5 holds. The rest of the claims therefore follow from that theorem, after proving the constraint qualification (3.13) for the equality claim.

From $(\Delta p, \Delta d) \in \mathcal{N}(\nabla G(p,d)) \cap T_{S_\epsilon}(p,d)$, we have by Lemma 3.4 and (3.15) that

$$\Delta d \in \mathcal{R}(B^*;A^*) \cap \mathcal{N}((I;0)(B^*;A^*)^\dagger) = \mathcal{R}(A^*), \text{ and } \Delta p \in \mathcal{N}(B_-) \cap \mathcal{N}(A). \tag{3.16}$$

When $\epsilon = 0$, Theorem 3.4 shows that under strict complementarity, $(\Delta p, \Delta d) \in T_{C_\epsilon}(p,d)$ reduces to $\Delta p = L(p)\eta$ and $\Delta d = -L(d)\eta$ for some $\eta \in \mathcal{J}$. Therefore the precedent of (3.13) implies

$$(L(p)\eta, -L(d)\eta) \in (\mathcal{N}(A) \cap \mathcal{N}(B_-)) \times \mathcal{R}(A^*). \tag{3.17}$$

Now non-degeneracy relative to $\mathcal{N}(B_-)$ forces $\eta = 0$, and then $(\Delta p, \Delta d) = 0$. □

The information in $D(\partial_\epsilon^{\mathrm{DC}} f_\nu)$ is not quite sufficient for our needs yet, so we extend it. More specifically, we let $\widehat{G}(p,d) \triangleq (G(p,d), p \circ d)$ for the G of Corollary 3.2, and consider

$$\widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu(y) \triangleq \{(z,q) \mid (y,z,q) \in \widehat{G}S_\epsilon\},$$
$$\widehat{U}_\epsilon(y|z,q)(\Delta y) \triangleq \bigcup_{(p,d)\in \widehat{G}_\epsilon^{-1}(y,z,q)} \{(\Delta z, \Delta q) \mid (\Delta y, \Delta z, \Delta q) \in \nabla \widehat{G}(p,d) T_{S_\epsilon}(p,d)\}.$$

The following assumption will be used frequently in what follows. Conditions ensuring the stated requirements will be further discussed in Section 3.4.5. Note that it may happen that $p \circ d \notin \mathcal{K}$.

Assumption 3.1. Let $(z,q) \in \widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu(y)$. Then $q = 0$ (resp. $q \in \mathcal{K} \setminus \{0\}$) and all $(p,d) \in \widehat{G}_\epsilon^{-1}(y,z,q)$ are strictly complementary and non-degenerate relative to $\mathcal{N}(B_-)$ (resp. $p, d \in \operatorname{int}\mathcal{K}$).

Lemma 3.5. *Suppose Assumption 3.1 holds and $\epsilon > 0$ (resp. $\epsilon = 0$). Then $(\Delta y, \Delta z, \Delta q) \in \nabla \widehat{G}(p,d) T_{S_\epsilon}(p,d)$ if and only if*

$$\operatorname{tr} \Delta q \leq \infty(\epsilon - \operatorname{tr} p \circ d) \quad (\textit{resp. } \Delta q = 0), \tag{3.18}$$

$$B^* \Delta y + A^* \Delta \lambda + \Delta d = 0, \tag{3.19}$$

$$A \Delta p = 0, \tag{3.20}$$

$$B_- \Delta p = \Delta z, \tag{3.21}$$

$$p \circ \Delta d + d \circ \Delta p = \Delta q. \tag{3.22}$$

Proof. Recalling (from the remark in Section 3.4.2) that non-degeneracy relative to $\mathcal{N}(B_-)$ is equivalent to that relative to $\mathcal{N}(B)$, we find that Assumption 3.1 guarantees the conditions for regularity in Lemma 3.4 and Theorem 3.4. Furthermore, the condition $(\Delta p, \Delta d) \in T_{\mathcal{K}}(p) \times T_{\mathcal{K}}(d)$ is redundant (by Theorem 3.4 when $\epsilon = 0$, and by $p, d \operatorname{int} \mathcal{K}$ when $\epsilon > 0$). Therefore

$$T_{S_\epsilon} = \{(\Delta p, \Delta d) \in \mathcal{J} \times \mathcal{J} \mid \text{(3.18)–(3.22) hold for some } (\Delta y, \Delta \lambda, \Delta z, \Delta q)\}.$$

Furthermore, we have

$$\nabla \widehat{G}(p,d) = \begin{pmatrix} 0 & -(I;0)(B^*; A^*)^\dagger \\ B_- & 0 \\ L(d) & L(p) \end{pmatrix}, \tag{3.23}$$

and therefore, employing (3.15), we find that $(\Delta y, \Delta z, \Delta q) \in \nabla \widehat{G}(p,d) T_{S_\epsilon}(p,d)$ forces (3.18)–(3.22) for this triple. □

Corollary 3.3. *$D(\widehat{\partial}^{\mathrm{DC}}_\epsilon f_\nu)(y|z,q) \supset \widehat{U}_\epsilon(y|z,q)$ with equality when Assumption 3.1 holds. For $\epsilon = 0$, $\widehat{\partial}^{\mathrm{DC}} f_\nu(y) = \partial^{\mathrm{DC}} f_\nu(y) \times \{0\}$.*

Proof. Noting that $q = p \circ d = 0$ when $\epsilon = 0$, takes care of the second claim.

From Lemma 3.5 we find that the constraint qualification (3.13) is equivalent to the linear system (3.19)–(3.22) for $(\Delta y, \Delta z, \Delta q) = 0$ implying $(\Delta p, \Delta d) = 0$. That is,

$$A \Delta p = 0,\ B_- \Delta p = 0,\ A^* \Delta \lambda + \Delta d = 0,\ L(p) \Delta d + L(d) \Delta p = 0 \tag{3.24}$$

has $(\Delta p, \Delta d, \Delta \lambda) = 0$ as the only solution

When $q = 0$, we proceed as in Corollary 3.2, applying Lemma 3.2.

When $q \in \mathcal{K} \setminus \{0\}$, we further tighten the uniqueness requirements by dropping $B_- \Delta p = 0$. Then the resulting system of equations is of a form familiar from linear programming on symmetric cones. Indeed, when furthermore $p \circ d \in \mathcal{K}$, the non-singularity of this system follows from assumptions (3.8)–(3.9) and standard results [Faybusovich, 1997b, Corollary 4.4].

With the constraint qualification now proved, we just apply Theorem 3.5. □

Remark 3.2. We may regard the q-component of $\widehat{\partial}^{\mathrm{DC}} f_\nu$ as indicating a specific "selection" $y \mapsto \{z \mid (z,q) \in \widehat{\partial}^{\mathrm{DC}}_\epsilon f_\nu(y)\}$ within $\partial^{\mathrm{DC}}_\epsilon f_\nu$, approximating the differences of subgradients of f and ν. In particular, the selections $q = (\epsilon/r)e$ give the gradients of barrier-approximations to f_ν; see Remark 3.1. So $\partial^{\mathrm{DC}}_\epsilon f_\nu$ is then a bundle with the information of the particular approximation lost, whereas $\widehat{\partial}^{\mathrm{DC}}_\epsilon f_\nu$ retains that information. $D(\widehat{\partial}^{\mathrm{DC}}_\epsilon f_\nu)(y|z,q)$ then combines the gradient of a selection with inter-selection differential information.

3.4.4 Solvability and regularity

In the interior point methods that we will develop in Chapter 4, it is of importance to know when we can solve $(0, \Delta q) \in D(\widehat{\partial}^{\mathrm{DC}}_\epsilon f_\nu)(y|z,q)(\Delta y)$ for Δy with fixed Δq, along with obtaining $(\Delta p, \Delta d)$. The following results study conditions towards that end.

Lemma 3.6. *Suppose Assumption 3.1 holds along with the following second order condition:* $0 \in D(\widehat{\partial}^{\mathrm{DC}}_\epsilon f_\nu)(y|z,q)(\Delta y)$ *implies* $\Delta y = 0$. *Then*

$$\Delta y \mapsto \{(\Delta z, \Delta q) \mid (\Delta y, \Delta z, \Delta q) \in \nabla \widehat{G}(p,d) T_{S_\epsilon}(p,d)\}$$

has full range for all $(p,d) \in \widehat{G}_\epsilon^{-1}(y,z,q)$. *Additionally the system* (3.19)–(3.22) *is solvable for* $(\Delta p, \Delta d, \Delta y, \Delta\lambda)$ *in a neighbourhood (in* $\mathcal{K} \times \mathcal{K}$*) of* $(p,d) \in \widehat{G}_\epsilon^{-1}(y,z,q)$.

Proof. By Lemma 3.5, the condition $0 \in D(\widehat{\partial}^{\mathrm{DC}}_\epsilon f_\nu)(y|z,q)(\Delta y)$ is equivalent to (3.19)–(3.22) with $(\Delta z, \Delta q) = 0$. Now, if also $\Delta y = 0$, the system further reduces into (3.24), and the proof of the constraint qualification (3.13) in Corollary 3.3 shows that there is no non-zero solution. Thus zero is the only solution of

$$\begin{pmatrix} A & & & \\ B_- & & & \\ & B^* & A^* & I \\ L(d) & & & L(p) \end{pmatrix} \begin{pmatrix} \Delta p \\ \Delta y \\ \Delta\lambda \\ \Delta d \end{pmatrix} = 0,$$

and the matrix has full rank. The same must hold in a neighbourhood of $(p,d) \in \widehat{G}_\epsilon^{-1}(y,z,q)$.

Finally, reverse application of Lemma 3.5 implies full range for $\Delta y \mapsto \{(\Delta z, \Delta q) \mid (\Delta y, \Delta z, \Delta q) \in \nabla \widehat{G}(p,d) \widehat{T}_{S_\epsilon}(p,d)\}$ for all $(p,d) \in \widehat{G}_\epsilon^{-1}(y,z,q)$. □

Remark 3.3. When f_ν is twice continuously differentiable at y, we have $D(\partial^{\mathrm{DC}} f_\nu)(y|z)(\Delta y) = \nabla^2 f_\nu(y)\Delta y$. Thus the second order condition reduces to non-singularity of the Hessian.

The picture is further revealed by considering the metric regularity of $\widehat{\partial}^{\mathrm{DC}}_\epsilon f_\nu$. A set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is said to be *metrically regular* at (y,z) if S^{-1} has the Aubin property at this point [Rockafellar and Wets, 1998, Theorem 9.43], i.e., if its graph is locally closed (i.e., has a closed neighbourhood of (y,z)) and there exist neighbourhoods $Y \ni y$ and $Z \ni z$, and a $\kappa > 0$ such that $S^{-1}(z'') \cap Y \subset S^{-1}(z') + \kappa\|z' - z''\|\mathbb{B}$ for all $z', z'' \in Z$, with $\mathbb{B}$ denoting the unit ball. The result

[Ibid., Theorem 9.40] states that under local closedness, metric regularity holds if $\mathcal{R}(\widehat{D}S(y|z)) = \mathbb{R}^m$, and it is equivalent to this property when S is graphically regular at (y, z).

Lemma 3.7. *Let*

$$W(y|z,q)(\Delta y) \triangleq \bigcap\nolimits_{(p,d)\in\widehat{G}_\epsilon^{-1}(y,z,q)} \{(\Delta z, \Delta q) \mid (\Delta y, \Delta z, \Delta q) \in \nabla\widehat{G}(p,d)\widehat{T}_{S_\epsilon}(p,d)\}.$$

Then the following regularity properties hold for $\widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu$ and $(z,q) \in \widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu(y)$.

(i) If $\widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu$ is graphically regular at (y,z,q), then it is metrically regular at this point if and only if $D(\widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu)(y|z,q)$ has full range.

(ii) Suppose graphical and metric regularity hold at (y,z,q) along with Assumption 3.1. Then $\widehat{U}_\epsilon(y|z,q)$ has full range.

(iii) $\widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu$ is metrically regular at (y,z,q) if $W(y|z,q)$ has full range.

Proof. (i) The graph $\widehat{G}S_\epsilon$ of $\widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu$ is locally closed: if a sequence $v_{[k]} = (y_{[k]}, z_{[k]}, q_{[k]}) \in \widehat{G}S_\epsilon$ has an accumulation point, it arises from bounded $(p_{[k]}, d_{[k]}) \in S_\epsilon$ as proved in Theorem 3.5, and S_ϵ is closed. Therefore the first claim follows directly from [Ibid., Theorem 9.40], quoted above.

(ii) Assumption 3.1 and Corollary 3.3 ensure $D(\widehat{\partial}_\epsilon^{\mathrm{DC}} f)(y|z,q) = \widehat{U}_\epsilon(y|z,q)$. Now apply (i).

(iii) By Theorem 3.3 $\widehat{D}(\widehat{\partial}_\epsilon^{\mathrm{DC}} f)(y|z,q)(\Delta y) \supset W(y|z,q)(\Delta y)$, and consequently the former has full range, the boundedness requirement (see Section 3.2.1) again proved as in Theorem 3.5. Now [Ibid., Theorem 9.40] yields metric regularity. □

3.4.5 Non-degeneracy

The following results ensure relative non-degeneracy, uniqueness, and Assumption 3.1. We often use

Assumption 3.2. $\mathcal{K} = \prod_{i=1}^{m_A} \mathcal{K}_i$ for symmetric cones $\mathcal{K}_i$ (in a Jordan algebra $\mathcal{J}_i$ of rank r_i), and $Ap = (\langle a_1', p_1\rangle, \ldots, \langle a_{m_A}', p_{m_A}\rangle)$ with $a_i' \in \operatorname{int}\mathcal{K}_i$ when $p = (p_1, \ldots, p_{m_A})$, $p_i \in \mathcal{K}_i$.

Lemma 3.8. *Suppose Assumption 3.2 holds and $b > 0$. Then* (3.8)–(3.10) *hold, and $(p,d) \in S_0$ and $L(d)\eta \in \mathcal{R}(A^*)$ imply $L(d)\eta = 0$.*

Proof. Assumptions (3.8)–(3.10) are immediate from the form of A. If $L(d)\eta = A^*\lambda$, we may assume $\lambda \geq 0$: by the independence of $L(d)$ on the sub-algebras corresponding to the $\mathcal{K}_i$, by negating components, we could find such a $\lambda' \geq 0$ and η' for which this holds. Therefore, unless $\lambda = 0$,

$$\langle a_i', p_i\rangle = b_i > 0 \tag{3.25}$$

implies $0 < \langle b, \lambda\rangle = \langle p, A^*\lambda\rangle = \langle p, L(d)\eta\rangle = \langle p \circ d, \eta\rangle = 0$. This is a contradiction, whence $L(d)\eta = 0$. □

Lemma 3.9. *Suppose that* $Wp = (\sum_i W_{1i}p_i, \ldots, \sum_i W_{Ni}p_i)$ *in addition to Assumption 3.2. Let* W' *denote* W *with those* W_{ji} *removed, for which* d_i *has rank* $r_i - 1$*. Likewise denote by* A' *the corresponding modification of* A*. Then* $L(p)\eta = 0$ *if*

$$\mathcal{N}(A') \cap \mathcal{N}(W') = \{0\}, \tag{3.26}$$

$(p, d) \in S_0$*,* $L(d)\eta = 0$*, and* $L(p)\eta \in \mathcal{N}(A) \cap \mathcal{N}(W)$*. Consequently, strict complementarity of* $(p, d) \in S_0$ *and* (3.26) *imply non-degeneracy relative to* $\mathcal{N}(W)$*.*

Proof. If d_i has rank $r_i - 1$, then p_i is proportional to a single primitive idempotent c complementary to d_i. This and $L(d_i)\eta_i = 0$ imply that $\eta_i \in \mathcal{R}(Q_c^*) = \mathcal{R}(Q_c) = \mathbb{R}c$ (as can be seen from the Q-decomposition of $L(p)$). Consequently $s_i \triangleq L(p_i)\eta_i \propto p_i$. But then $s_i \in \pm\mathcal{K}_i$, which is in contradiction to $\langle a_i', s_i\rangle = 0$ unless $s_i = 0$, since $a_i' \in \operatorname{int}\mathcal{K}_i$. Therefore $L(p_i)\eta_i = 0$, and we may consequently remove the corresponding terms from the equations $WL(p)\eta = 0$ and $AL(p)\eta = 0$. The resulting equation has no non-zero solution when $\mathcal{N}(A') \cap \mathcal{N}(W') = \{0\}$.

As for the final claim, Lemma 3.8 reduces the non-degeneracy requirement relative to $\mathcal{N}(W)$ into $(L(d)\eta, L(p)\eta) \in \{0\} \times (\mathcal{N}(A) \cap \mathcal{N}(W))$ implying $\eta = 0$. Since $L(d + p)$ is invertible when p and d are strictly complementary, it suffices to show that $L(p)\eta = 0$. The first part of this lemma did that. □

Corollary 3.4. *Suppose each* $\mathcal{J}_i$ *has rank* $r_i = 2$ *(i.e.,* $\mathcal{K}_i$ *is isomorphic to the second order cone), and* $\mathcal{N}(W_{ji}) \cap \mathcal{N}(\langle a_i', \cdot\rangle) = \{0\}$*. Then strictly complementary* (p, d) *are non-degenerate relative to* $\mathcal{N}(W)$ *when for each* $j = 1, \ldots, N$*, at most one* $d_i = 0$ *with* $W_{ji} \neq 0$*.*

Proof. When $d_i \neq 0$, $p_i \neq 0$ is proportional to a single primitive idempotent. Consequently W' has just one non-zero W_{ji} on each row. But by assumption $\mathcal{N}(W_{ji}) \cap \mathcal{N}(\langle a_i', \cdot\rangle) = \{0\}$, so (3.26) holds. Therefore Lemma 3.9 provides non-degeneracy. □

The following results prove and simplify Assumption 3.1 through uniqueness.

Lemma 3.10. *Suppose* $(p, d) \in \widehat{G}_0^{-1}(y, z, 0)$ *is strictly complementary and non-degenerate relative to* $\mathcal{N}(B_-)$*. Then it is unique. In particular, Assumption 3.1 holds.*

Proof. Suppose $(p + \Delta p, d + \Delta d) \in \widehat{G}_0^{-1}(y, z, 0)$. Then $\Delta d \in \mathcal{R}(A^*)$ and $\Delta p \in \mathcal{N}(A) \cap \mathcal{N}(B_-)$. Consequently $\operatorname{tr}\Delta p \circ \Delta d = 0$. As $p \circ d = (p + \Delta p) \circ (d + \Delta d) = 0$, taking the trace we then find that $\operatorname{tr}(p \circ \Delta d + d \circ \Delta p) = 0$. This says that $\operatorname{tr}(p + \alpha\Delta p) \circ (d + \alpha\Delta d) = 0$ for all $\alpha \in [0, 1]$. Because $p + \alpha\Delta p, d + \alpha\Delta d \in \mathcal{K}$ by convexity, we find that $(p + \alpha\Delta p) \circ (d + \alpha\Delta d) = 0$. Differentiating $(p + \alpha\Delta p) \circ (d + \alpha\Delta d)$ at $\alpha = 0$, we find $p \circ \Delta d + d \circ \Delta p = 0$. Now strict complementarity and Lemma 3.2(i) imply $(\Delta p, \Delta d) = (L(p)\eta, -L(d)\eta)$ for some $\eta \in \mathcal{J}$. By non-degeneracy $\eta = 0$. Therefore (p, d) is unique. □

Lemma 3.11. *Suppose* $p, d \in \mathcal{K}$ *and* $q = p \circ d \in \operatorname{int}\mathcal{K}$*. Then* $p, d \in \operatorname{int}\mathcal{K}$*, so Assumption 3.1 holds.*

Proof. If $d \in \operatorname{bd}\mathcal{K}$, there is a $v \in \mathcal{K} \setminus \{0\}$ such that $v \circ d = 0$ (as this is equivalent to $\langle v, d\rangle = 0$). Now, $\langle v, q\rangle = \langle v, p \circ d\rangle = \langle v \circ d, p\rangle = 0$, in contradiction to $q \in \operatorname{int}\mathcal{K}$. The case $p \in \operatorname{bd}\mathcal{K}$ is analogous. □

Lemma 3.12. *If $q \in \operatorname{int}\mathcal{K}$ and Assumption 3.2 holds, then there is at most one $(p, d) \in \widehat{G}_\infty^{-1}(y, z, q)$.*

Proof. Let λ be such that $p, d \in \mathcal{K}$ when $B^*y + A^*\lambda + d + c = 0$ and p is defined through $p \circ d = q$. Due to Lemma 3.11, actually $p, d \in \operatorname{int}\mathcal{K}$. Choose a direction $\Delta\lambda$, and differentiate these equations to find $A^*\Delta\lambda + \Delta d = 0$, $L(p)\Delta d + L(d)\Delta p = 0$. Solving for Δp and taking the inner product with Δd yields

$$\langle \Delta d, \Delta p\rangle = -\langle \Delta d, L(d)^{-1}L(p)\Delta d\rangle = -\langle \Delta d, (L(d)^{-1}L(p) + L(p)L(d)^{-1})\Delta d\rangle/2.$$

The operator on the right may be expanded as $L(d)^{-1}\big(L(p)L(d) + L(d)L(p)\big)L(d)^{-1}$. The middle term is positive-definite according to [Faybusovich, 1997b, proof of Corollary 4.4]. Therefore the entire operator is positive-definite, and consequently $-\langle \Delta\lambda, A\Delta p\rangle = \langle \Delta d, \Delta p\rangle < 0$ if $\Delta\lambda \neq 0$.

Now, since A is independent in each $\mathcal{K}_i$ under Assumption 3.2, $\langle a_i', \Delta p_i\rangle > 0$ whenever $\Delta\lambda_i > 0$. Thus $\langle a_i', p_i\rangle$ is an increasing function of λ_i. Consequently $\langle a_i', p_i\rangle = b_i$ has a unique solution. □

Example 3.6 (Sums of Euclidean norms)**.** Suppose $f(y) = \sum_{i=1}^n \|y - c_i\|$ and $c_i \neq c_j$ for $i \neq j$. Strict complementarity implies non-degeneracy and Assumption 3.1, because at most one term is non-differentiable at a single point, with corresponding $p_i \in \operatorname{int}\mathcal{K}_i$, and $\langle a_i', p_i\rangle = p_i^0$, $W_{ji}p_i = \bar{p}_i$. Thus the linear independence condition holds. Similar results hold for more complex combinations of norms; cf. also [Qi et al., 2002, Section 3].

3.4.6 Scaling

The following scaling invariance of the presentation of f, and by extension f_v, holds with respect to the automorphisms of the cone $\mathcal{K}$.

Lemma 3.13. *Let f have the form* (3.7)*, and let $v \in \operatorname{int}\mathcal{K}$. Define*

$$\tilde{f}(y) \triangleq \sup\{\langle \underset{\sim}{B}^*y + \underset{\sim}{c}, \tilde{p}\rangle \mid \tilde{p} \in \mathcal{K}, \underset{\sim}{A}\tilde{p} = b\}$$

with $\underset{\sim}{B} \triangleq BQ_v^{-1}$, $\underset{\sim}{A} \triangleq AQ_v^{-1}$, and $\underset{\sim}{c} = Q_v^{-1}c$. Then $\tilde{f} = f$ with $\tilde{p} = Q_v p$ producing the same value. In the representation of $\partial_\epsilon f$, same result is produced when furthermore $\underset{\sim}{d} = Q_v^{-1}d$. This scaling invariance extends to $\partial_\epsilon^{\mathrm{DC}} f_v$ in the obvious way.

Proof. Firstly note that assumptions (3.8)–(3.9) as well as the property $\mathcal{N}(\underset{\sim}{B}^*; \underset{\sim}{A}^*) = \{0\}$ continue to hold after scaling, so $\tilde{f}$ has the required form (3.7). Now the claims follow in a straightforward manner from Q_v being a bijection in $\mathcal{K}$. □

Note, however, that the q of $(z, q) \in \widehat{\partial}_\epsilon^{\mathrm{DC}} f_v(y)$ generally depends on the scaling. In the special case of the "central selection" $q = \mu e$, it is unaffected, as seen from [Schmieta and Alizadeh, 2003, Lemma 28] for $\mu > 0$ and from Lemma 3.14 below for $\mu = 0$.

Lemma 3.14. *Scaling as above preserves (strict) complementarity.*

Proof. When p and d are strictly complementary, $d \in \operatorname{ri} N_{\mathcal{K}}(p)$, as follows from Lemma 3.1. But now

$$\begin{aligned} N_{\mathcal{K}}(p) &= -\{s \in \mathcal{K} \mid \langle p, s\rangle = 0\} \\ &= -\{s \in \mathcal{K} \mid \langle Q_v p, Q_v^{-1} s\rangle = 0\} = -Q_v\{\underset{\sim}{s} \in \mathcal{K} \mid \langle \widetilde{p}, \underset{\sim}{s}\rangle = 0\}, \end{aligned}$$

so that $N_{\mathcal{K}}(\widetilde{p}) = Q_v^{-1} N_{\mathcal{K}}(p)$. As linear transformations map relative interior onto relative interior [Rockafellar, 1972, Theorem 6.6], $\underset{\sim}{d} = Q_v^{-1} d \in \operatorname{ri} N_{\mathcal{K}}(\widetilde{p})$.

When the complementarity is non-strict, the same argument holds without the relative interior taken. □

Relative non-degeneracy and Assumption 3.1 are also preserved by scaling under the conditions of Section 3.4.5:

Lemma 3.15. *Assumption 3.2, the representation of Lemma 3.9, and* (3.26) *are preserved by scaling.*

Proof. The transformation Q_v decomposes into a product $(Q_{v_1}, \ldots, Q_{v_{ma}})$, so that Q_v maps each $\mathcal{K}_i$ independently onto itself and $\underset{\sim}{a}_i' = Q_{v_i}^{-1} a_i' \in \operatorname{int} \mathcal{K}_i$. Likewise $\underset{\sim}{W}$ is composed of $\underset{\sim}{W}_{ij} = W_{ij} Q_{v_i}^{-1}$. Thus the special forms of A and W in Assumption 3.2 and Lemma 3.9 are preserved by scaling.

By the proof of Lemma 3.14, scaling preserves the rank of d_i. Now note that $\mathcal{N}(\underset{\sim}{A}') = Q_{v'} \mathcal{N}(A')$ (for suitably modified v') and likewise for $\underset{\sim}{W}'$. Therefore $\mathcal{N}(\underset{\sim}{A}') \cap \mathcal{N}(\underset{\sim}{W}') = \{0\}$ if and only if $\mathcal{N}(A') \cap \mathcal{N}(W') = \{0\}$. The assumptions of Lemma 3.9 thus continue to hold after scaling. □

4 PRIMAL-DUAL INTERIOR POINT METHODS FOR DIFF-CONVEX PROBLEMS ON SYMMETRIC CONES

4.1 Introduction

As seen in Chapter 3, the optimality conditions (3.2) for convex functions of the form (3.1) belong to the same class as those for linear programs on symmetric cones and, as already discussed in Chapter 1 very efficient algorithms exist for approximately solving such equations; cf., e.g., [Nesterov and Todd, 1997; Schmieta and Alizadeh, 2003, 2001; Faybusovich, 1997a,b; Alizadeh and Goldfarb, 2003; Monteiro and Tsuchiya, 2000] in more general cases, and [Andersen et al., 2000; Xue and Ye, 1997; Qi et al., 2002] in the special case of Euclidean norms, various sums of which are included in the class of functions of the form (3.1).

As it has also turned out to be, those approximate solutions in the present case correspond to $0 \in \partial_\epsilon f(y)$ with some additional conditions on choice of selection within the subdifferential. In the present chapter, we intend to extend these methods to solving $0 \in \partial_\epsilon^{\mathrm{DC}} f_\nu(y)$, with the consequent central path conditions

$$B^*y + A^*\lambda + d + c = 0,\ Ap = b, B_-p = 0,\ p \circ d = (\epsilon/r)e,\ p, d \in \mathcal{K}. \tag{4.1}$$

The extension of the interior point methods faces the problem that the linearised version of (4.1) may become singular, something that does not occur in the convex case under rather mild assumptions. As the first topic of Section 4.2, we therefore analyse such singularities through the graphical (second-order) differentials derived in Chapter 3, before extending the aforementioned methods in the same section. It will be seen that our extension still bears good convergence properties near a point satisfying rather standard second order (optimality) conditions.

Aside from the general literature on interior point methods (see, e.g., Forsgren et al. [2002] and references therein), and the already-cited papers on linear programs over symmetric cones, the work of Yamashita and Yabe [2005] bears

close relationship to ours, in extending the Jordan-algebraic approach. There nonlinear programs are considered, however only over second-order cones rather than general symmetric cones, and with C^2 assumptions. The analysis is also vastly different from ours, based on merit functions.

As our extension is, however, not globally convergent due to the above-mentioned singularities, we next study globalisation strategies in Section 4.3. Our approach is that of a filter method, following the line of research initiated by Fletcher and Leyffer [2002].

Filter methods, which we will introduce in further detail in Section 4.3, crucially depend on so-called restoration methods that restore feasibility after the main filter method – which will presently be a variation of the interior point method of Section 4.2 – runs into trouble. Our next topic is therefore to derive and analyse one in Section 4.4, based on the simple idea of sequential convex programming (SCP).

We finish this chapter with a discussion of practical aspects and experience in Section 4.5.

4.2 A primal-dual interior point method

4.2.1 On interior point methods for the convex case

Suppose we are given a point $0 \in \partial_\epsilon f(y)$. To minimise f, we want to reduce ϵ, while at the same time keeping the constraint $0 \in \partial_\epsilon f(y)$. Thus we want to choose a direction Δy such that $0 \in D(\partial_\epsilon f)(y|0)(\Delta y)$ and ϵ can be reduced afterwards. If $(y, 0) \in \operatorname{int} \operatorname{Graph} \partial_\epsilon f(y)$, any direction satisfies this. When we additionally want to be moving towards a "central selection" from a selection q with $(0, q) \in \widehat{\partial}_\epsilon f(y)$, we require that $(0, \Delta q) \in D(\widehat{\partial}_\epsilon f)(y|0, q)(\Delta y)$ for $\Delta q \triangleq \sigma\mu e - q$, $\mu = \mu(q) \triangleq \operatorname{tr} q/r$, and a chosen $\sigma \in (0, 1)$. We may think of Δq consisting of a "tangential step" $(\sigma - 1)\mu e$ aiming to reduce μ or ϵ, and a "normal step" $\mu e - q$ aiming to move closer to the central selection for $\partial_{r\mu} f$.

Suppose furthermore that we have $(p, d) \in \widehat{G}_\epsilon^{-1}(y, 0, q)$, and want to make our movement in the neighbourhood of (p, d). Then $q = p \circ d$, and by the proof of Lemma 3.5, we arrive from $(\Delta y, 0, \sigma\mu e - q) \in \nabla \widehat{G}(p, d) T_{S_\epsilon(p,d)}$ into the system

$$\begin{gathered} A\Delta p = 0, B\Delta p = 0, \\ B^*\Delta y + A^*\Delta\lambda + \Delta d = 0, \\ p \circ \Delta d + d \circ \Delta p = \sigma\mu e - p \circ d, \\ \Delta p \in T_{\mathcal{K}}(p), \Delta d \in T_{\mathcal{K}}(d). \end{gathered}$$

When $p, d \in \operatorname{int} \mathcal{K}$ and $p \circ d \in \mathcal{K}$, the linear system is solvable. By iterating steps in directions found this way after suitable scaling and step length selection, we get the usual primal-dual interior point method for linear programs on symmetric cones [Nesterov and Todd, 1997; Schmieta and Alizadeh, 2003, 2001; Faybusovich, 1997a,b; Alizadeh and Goldfarb, 2003; Monteiro and Tsuchiya, 2000].

Whereas typically the "interior" refers to the interior of a constraint set, and the above system of equations have been derived through either the use of barrier functions, or by perturbation of the KKT conditions, here the conditions have been derived through subdifferential analysis, and we can alternatively consider to be moving in the interior of $\partial_\epsilon f$ and even the set $\widehat{G}S_\infty = \operatorname{Graph} \widehat{\partial}_\infty f$, while maintaining the ϵ-optimality constraint $0 \in \partial_\epsilon f(y)$, reducing ϵ by a constant factor at each iteration. Additionally, we try to stay close to a "central selection" $p \circ d = \mu e$, corresponding to the differential of a smoothing of f by a barrier function.

4.2.2 Solvability in the diff-convex case

Our objective is now analogous to the convex case: given $(0, q) \in \widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu(y)$ and $(p, d) \in \widehat{G}_\epsilon^{-1}(y, 0, q)$, we try to solve $(0, \Delta q) \in D(\widehat{\partial}_\epsilon^{\mathrm{DC}} f)(y|0, q)(\Delta y)$ near (p, d). When $p, d \in \operatorname{int} \mathcal{K}$ and $\Delta q = \sigma\mu e - p \circ d$, the resulting set of equations may then according to Lemma 3.5 be written

$$A\Delta p = 0, B_- \Delta p = 0, \tag{4.2}$$

$$B^* \Delta y + A^* \Delta\lambda + \Delta d = 0, \tag{4.3}$$

$$p \circ \Delta d + d \circ \Delta p = \sigma\mu e - p \circ d. \tag{4.4}$$

This differs from the convex case by the use of B_- instead of B in the condition for Δp. Consequently, we run into the following two problems in a direct generalisation of the methods for convex problems: (a) we may have $\langle \Delta p, \Delta d \rangle \neq 0$, and (b) the system may not have a solution for any specific value of Δq. Therefore other strategies are needed for global convergence. But let us first analyse how far a direct generalisation goes, and its convergence properties.

According to the results of Section 3.4.4 and Lemma 3.6 in particular, the system (4.2)–(4.4) can be solved at least locally in the neighbourhood of a point y arising from relatively non-degenerate and strictly complementary (p, d), and where $0 \in D(\widehat{\partial}^{\mathrm{DC}} f_\nu)(y|0)(\Delta y)$ implies $\Delta y = 0$. Furthermore, this second order condition reduces to non-singularity of the Hessian when f_ν is twice continuously differentiable.

Likewise, by the same lemma, the system (4.2)–(4.4) is solvable near nicely-behaving selections of $\widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu$. Also, since central selections $q = \mu e \in \mathcal{K}$, $\mu > 0$, are unaffected by scaling as remarked in Section 3.4.6, the same applies to scaled representation of f near central selections.

Study of metric regularity offers some further insight. Statement (i) of Lemma 3.7 says that $D(\widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu)(y|z, q)$ can behave badly when $\widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu$ is not metrically regular. In particular (assuming $\epsilon > \operatorname{tr} q$, or $\epsilon = 0$ and working with $\partial^{\mathrm{DC}} f_\nu$), if there exists (z', q') in each neighbourhood of (z, q) such that $(z', q') \notin \mathcal{R}(\widehat{\partial}_\epsilon^{\mathrm{DC}} f)$, then we may not be able to solve (4.2)–(4.4). This happens in particular when $\operatorname{Graph} \partial_\epsilon^{\mathrm{DC}} f_\nu$ locally evades the $z = 0$ plane as ϵ shrinks. Another way of metric regularity failing is that for some q' in each neighbourhood of q, the selection $y \mapsto \{z \mid (z, q') \in \widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu(y)\}$ fails to be metrically regular at (y, z),

and consequently has a singular (or otherwise poorly-behaved graphical second order) differential.

On the other hand, by Lemma 3.7(ii), metrical and graphical regularity imply for every $(\Delta z, \Delta q)$, the existence of some $(p, d) \in \widehat{G}_\epsilon^{-1}(y, z, q)$ such that (4.2)–(4.4) has a solution. Under Assumption 3.2, by Lemma 3.12, this (p, d) is unique, so the system is invertible. Conversely, Lemma 3.7(iii) says that $\widehat{\partial}_\epsilon^{\mathrm{DC}} f_\nu$ actually is metrically (and graphically) regular at (y, z, q) provided that $(p, d) \in \widehat{G}_\epsilon^{-1}(y, z, q)$ is unique, the matrix of the system (4.2)–(4.4) has full range, and Assumption 3.1 holds (such as when $q \in \operatorname{int} \mathcal{K}$).

4.2.3 Neighbourhoods

Let $P_e^\perp q \triangleq q - \langle e, q\rangle e/r$ be the projection of q to the subspace orthogonal to e. If the spectrum of q is $\{\zeta_i(q)\}$, then by the e-sum property of Jordan frames, the spectrum of $P_e^\perp q$ is $\{\zeta_i(q) - \mu(q)\}$ with $\mu(q) \triangleq \sum_j \zeta_j(q)/r = \operatorname{tr} q/r$. Now, define the distance functions

$$d_\bullet(p, d) \triangleq \|P_e^\perp Q_p^{1/2} d\|_\bullet \quad \text{and} \quad d_\bullet^*(p, d) \triangleq \|P_e^\perp (p \circ d)\|_\bullet,$$

with $\bullet \in \{F, 2, -\infty\}$ and, abusing norm notation for the sake of convenience, $\|s\|_{-\infty} \triangleq -\min_i \zeta_i(s)$. For $P_e^\perp q$ we then get $\|P_e^\perp q\|_{-\infty} = \mu(q) - \min \zeta_i(q)$, $\|P_e^\perp q\|_F = \sqrt{\sum_i (\zeta_i(q) - \mu(q))^2}$, and $\|P_e^\perp q\|_2 = \max_i |\zeta_i(q) - \mu(q)|$.

When $p, d \in \operatorname{int} \mathcal{K}$, we know from the effects of $P_e^\perp$ on the spectrum and [Schmieta and Alizadeh, 2003, Proposition 21 and Lemma 30] that $d_\bullet(d, p) = d_\bullet(p, d) \leq d_\bullet^*(p, d) = d_\bullet^*(d, p)$ for $p, d \in \operatorname{int} \mathcal{K}$. When p and d operator-commute, equality holds as then $p \circ d = Q_p^{1/2} d$.

Now, let $\gamma \in (0, 1)$, and for $\bullet \in \{F, 2, -\infty\}$ define the corresponding short, semi-long, and long-step neighbourhoods of $\mathcal{K} \times \mathcal{K}$ as

$$\begin{aligned} \mathcal{C}_\bullet(\gamma) &\triangleq \{(p, d) \in \operatorname{int} \mathcal{K} \times \operatorname{int} \mathcal{K} \mid d_\bullet(p, d) \leq \gamma \mu(p \circ d)\} \quad \text{and} \\ \mathcal{C}_\bullet^*(\gamma) &\triangleq \{(p, d) \in \operatorname{int} \mathcal{K} \times \operatorname{int} \mathcal{K} \mid d_\bullet^*(p, d) \leq \gamma \mu(p \circ d)\}, \end{aligned}$$

We then have $\mathcal{C}_\bullet^*(\gamma) \subset \mathcal{C}_\bullet(\gamma)$, as well as $\mathcal{C}_F(\gamma) \subset \mathcal{C}_2(\gamma) \subset \mathcal{C}_{-\infty}(\gamma)$, and likewise for the starred neighbourhoods. The unstarred neighbourhoods are scaling-invariant, i.e., $(p, d) \in \mathcal{C}_\bullet(\gamma)$ implies $(\widetilde{p}, \underset{\sim}{d}) = (Q_v p, Q_v^{-1} d) \in \mathcal{C}_\bullet(\gamma)$ for $v \in \operatorname{int} \mathcal{K}$ [Schmieta and Alizadeh, 2003, Proposition 29]. Furthermore, a scaling that results in operator-commutative $(\widetilde{p}, \underset{\sim}{d})$ ensures that $(\widetilde{p}, \underset{\sim}{d}) \in \mathcal{C}_\bullet^*(\gamma)$ for $(p, d) \in \mathcal{C}_\bullet(\gamma)$.

In the method we keep (p, d) in an appropriate γ-neighbourhood to ensure desirable properties, such as $p \circ d \in \operatorname{int} \mathcal{K}$ (cf. Lemma 3.11).

4.2.4 Rate of convergence

We now provide some rate of convergence properties, assuming we have a solution $(\Delta p, \Delta d)$ of (4.2)–(4.4). The proofs here follow the outline of [Schmieta and Alizadeh, 2003, Section 3], generalising where necessary to accommodate $\langle \Delta p, \Delta d\rangle \neq 0$, and also to rely less on operator-commutativity. We note that our

analysis does not actually depend on the exact form of the linear equations (4.2)–(4.3). These conditions merely act as source of proximity to singularities for the whole system, and therefore the analysis could easily be applied to other linear systems sharing (4.4), arising from optimality conditions for more general classes of problems.

So, let us set

$$\begin{aligned} p(\alpha) &\triangleq p + \alpha\Delta p, \quad d(\alpha) \triangleq p + \alpha\Delta p, \\ \mu(\alpha) &\triangleq \operatorname{tr} p(\alpha) \circ d(\alpha)/r. \end{aligned} \tag{4.5}$$

Then, denoting $\Delta \triangleq \Delta p \circ \Delta d$,

$$\begin{aligned} r\mu(\alpha) &= \operatorname{tr} p \circ d + \alpha \operatorname{tr}(p \circ \Delta d + d \circ \Delta p) + \alpha^2 \operatorname{tr} \Delta p \circ \Delta d \\ &= r\mu + \alpha(\sigma - 1)r\mu + \alpha^2 \operatorname{tr} \Delta \\ &= (1 - \alpha)r\mu + \alpha\sigma r\mu + \alpha^2 \operatorname{tr} \Delta. \end{aligned} \tag{4.6}$$

The linear constraints of $(p(\alpha), d(\alpha)) \in S_{r\mu(\alpha)}$ obviously automatically continue to hold for any α. The next lemma bounds the non-linear constraints.

Lemma 4.1. *If $(p, d) \in \mathcal{C}^*_\bullet(\gamma)$ for some $\bullet \in \{F, 2, -\infty\}$, then $(p(\alpha), d(\alpha)) \in \mathcal{C}^*_\bullet(\gamma) \cup C_0$ for $\alpha \in [0, \bar{\alpha}]$, where*

$$\bar{\alpha} \triangleq \begin{cases} \sigma/\kappa, & \kappa \geq \sigma, \\ 1/(1 - \sigma/2), & \kappa = 0, \\ \sqrt{(1 - \sigma/2)^2/\kappa^2 + 2/\kappa} - (1 - \sigma/2)/\kappa, & 0 \neq \kappa \in (-(1 - \sigma/2)^2/2, \sigma), \\ \infty, & \textit{otherwise}, \end{cases} \tag{4.7}$$

and $\kappa \triangleq (\|P_e^\perp \Delta\|_F - \gamma \operatorname{tr} \Delta/r)/(\gamma\mu)$. When $\kappa < \sigma$, then $\bar{\alpha} > 1$.

Proof. It suffices to prove that for $\alpha \in (0, \bar{\alpha})$, $\|P_e^\perp(p(\alpha) \circ d(\alpha))\|_\bullet < \gamma\mu(\alpha)$. For, as follows from the relationships presented in Section 4.2.3, then the same holds for $\bullet = -\infty$, and consequently

$$(1 - \gamma)\mu(\alpha) < \min_i \zeta_i(p(\alpha) \circ d(\alpha)) \leq \min_i \zeta_i(Q^{1/2}_{p(\alpha)} d(\alpha)),$$

where the second inequality is proved in [Schmieta and Alizadeh, 2003, Lemma 30], and applies when $p(\alpha) \in \operatorname{int} \mathcal{K}$. But then, taking the power of r on both sides, we get

$$((1 - \gamma)\mu(\alpha))^r < \det(Q^{1/2}_{p(\alpha)} d(\alpha)) = \det(p(\alpha)) \det(d(\alpha)),$$

applying [Faraut and Korányi, 1994, Proposition III.4.2] on subalgebras for the equality. Now, by the continuity of the involved quantities in α, this condition would be violated if at some point either $p(\alpha)$ or $d(\alpha)$ reached $\operatorname{bd} \mathcal{K}$ while still $\mu(\alpha) > 0$. But if $\mu(\alpha) = 0$, we must also have $\|P_e^\perp(p(\alpha) \circ d(\alpha))\|_\bullet = 0$, whence $\alpha = \bar{\alpha}$. Thus $(p(\alpha), d(\alpha)) \in C_0$, and we have a solution to the problem.

We have

$$\begin{aligned} P_e^\perp(p(\alpha)\circ d(\alpha)) &= P_e^\perp(p\circ d) + \alpha P_e^\perp(p\circ \Delta d + d\circ \Delta p) + \alpha^2 P_e^\perp(\Delta d\circ \Delta p) \\ &= P_e^\perp(p\circ d) + \alpha P_e^\perp(\sigma\mu e - p\circ d) + \alpha^2 P_e^\perp \Delta \\ &= (1-\alpha)P_e^\perp(p\circ d) + \alpha^2 P_e^\perp \Delta. \end{aligned}$$

To approximate the norm, for $\bullet = F$ we can use the triangle inequality, whereas for $\bullet = 2, -\infty$, we apply [Schmieta and Alizadeh, 2003, Lemma 14], which states that for $x, y \in \mathcal{J}$, $-\min\zeta_i(x+y) \le -\min\zeta_i(x) + \|y\|_F$, and $\max\zeta_i(x+y) \le \max\zeta_i(x) + \|y\|_F$. Therefore, for all $\bullet \in \{F, 2, -\infty\}$, we have the approximation

$$\begin{aligned} \|P_e^\perp(p(\alpha)\circ d(\alpha))\|_\bullet &\le |1-\alpha|\|P_e^\perp Q_p^{1/2} d\|_\bullet + \alpha^2\gamma\|P_e^\perp\Delta\|_F \\ &\le |1-\alpha|\gamma\mu + \alpha^2\|P_e^\perp\Delta\|_F. \end{aligned}$$

Comparing this approximation against $\mu(\alpha)$ from (4.6), we get that

$$\|P_e^\perp(p(\alpha)\circ d(\alpha))\| \le \gamma\mu(\alpha)$$

if

$$\alpha^2\|P_e^\perp\Delta\|_F \le (1-\alpha-|1-\alpha|+\alpha\sigma)\gamma\mu + \gamma\alpha^2 \operatorname{tr}\Delta/r,$$

i.e., $\alpha^2\kappa \le (1-\alpha-|1-\alpha|+\alpha\sigma)$.

Suppose we have equality at $0 < \alpha \le 1$. Then $\kappa \ge \sigma$, and we get the bound in (4.7). On the other hand, if $\kappa < \sigma$, the inequality holds strictly for all $\alpha \in (0,1]$. So equality is reached at $\alpha > 1$, and we get the bound in (4.7) by solving the quadratic equation $\alpha^2\kappa - 2 + \alpha(2-\sigma) = 0$. When $\kappa \ne 0$, there are potentially two solutions,

$$\alpha = \frac{-(1-\sigma/2) \pm \sqrt{(1-\sigma/2)^2 + 2\kappa}}{\kappa},$$

but the bound in (4.7) is the one we want. This follows for $\kappa > 0$, because the other solution is negative. For $\kappa < 0$ this follows from observing that a quadratic function with a negative quadratic term, which is also negative and increasing at $\alpha = 0$, has only positive roots, if any. Therefore the smaller root, if any, gives the bound, and otherwise it is infinite. Solving for the term under the square root to equal zero gives the lower bound for the applicability of the expression in (4.7). □

Suppose $\operatorname{tr}\Delta > 0$. Then, minimising $\mu(\alpha)$ over $\alpha \ge 0$, we get $\sigma\mu = 2\check{\alpha}\operatorname{tr}\Delta$, or $\check{\alpha} \triangleq (1-\sigma)/(2\check{\kappa})$ with $\check{\kappa} \triangleq \operatorname{tr}\Delta/(r\mu)$. For convenience, we set $\check{\alpha} = \infty$ when $\operatorname{tr}\Delta \le 0$.

Lemma 4.2. *Suppose the conditions of Lemma 4.1 hold. Let* $\hat{\alpha} \triangleq \min\{\bar{\alpha}, \check{\alpha}\}$. *Then*

$$\delta \triangleq 1 - \mu(\hat{\alpha})/\mu \ge (1-\sigma)\hat{\alpha}/2. \tag{4.8}$$

Proof. When $\operatorname{tr}\Delta > 0$, $\check{\alpha} \ge \alpha$ is equivalent to $\check{\kappa}\alpha \le (1-\sigma)/2$. Then we find from (4.6) that

$$\begin{aligned} \mu(\alpha)/\mu - 1 &= (\sigma-1)\alpha + \alpha^2\check{\kappa} \\ &\le (\sigma-1)\alpha + (1/2)(1-\sigma)\alpha = (1/2)(\sigma-1)\alpha. \end{aligned}$$

When $\operatorname{tr}\Delta \leq 0$, the same result continues to hold because $\alpha^2\check{\kappa} \leq 0$ may be dropped, and $\sigma - 1 < 0$. Therefore the claim holds when $\check{\alpha} \geq \bar{\alpha}$.

When $\check{\alpha} \leq \bar{\alpha}$, we get that $\mu(\check{\alpha})/\mu - 1 = (\sigma - 1)\check{\alpha} + (1 - \sigma)\check{\alpha}/2$, which gives the desired result. □

Therefore, to obtain fast decrease in μ, it suffices to bound $\hat{\alpha}$ from below. For, given a lower bound $\hat{\delta} \leq \delta$, a standard argument[1] shows that $\hat{\delta}^{-1}\log\tau^{-1}$ steps are sufficient to ensure that $\mu \leq \tau\overline{\mu}$ for an initial $\overline{\mu} > 0$ and desired decrease factor $\tau \in (0, 1)$.

If $\kappa < \sigma$, then $\bar{\alpha} > 1$ from Lemma 4.1. Therefore in this case, it suffices to have a bound for $\check{\alpha}$ from below. Consequently, it suffices to bound both κ and $\check{\kappa}$ from above. Let us see how far that can be done.

Lemma 4.3. *Let $u, v \in \mathcal{J}$ and let H_u and H_v be invertible linear operators on $\mathcal{J}$, with the induced norm $\|H\|_F \triangleq \max_{x\neq 0}\|Hx\|_F/\|x\|_F$. Then*

$$\|u\|_F\|v\|_F \leq \frac{1}{2}\|H_u^{-1}\|_F\|H_v^{-1}\|_F\big(\|H_u u\|^2 + \|H_v v\|^2\big).$$

Proof. We have $\|u\|_F = \|H_u^{-1}H_u u\| \leq \|H_u^{-1}\|_F\|H_u u\|_F$ and likewise for v. Now apply the inequality $2ab \leq a^2 + b^2$. □

Lemma 4.4. *Suppose $p, d, q = p \circ d \in \operatorname{int}\mathcal{K}$, and that (4.4) holds. Suppose H_0 is an invertible linear operator in $\mathcal{J}$ that satisfies $H_0 q = q^{1/2}$ and $H_0 e = q^{-1/2}$. Let $H_d \triangleq H_0 L(p)$ and $H_p \triangleq H_0 L(d)$. Then $\|H_d\Delta d\|_F^2 + \|H_p\Delta p\|_F^2 = \theta - 2\langle H_p\Delta d, H_d\Delta d\rangle$ with*

$$\theta \triangleq \theta(q, \sigma) \triangleq \sum_{i=1}^{r}\frac{(\sigma\mu(q) - \zeta_i(q))^2}{\zeta_i(q)}.$$

Proof. Multiplying (4.4) from the left by H_0, we get

$$H_d\Delta d + H_p\Delta p = H_0(\sigma\mu e - p \circ d) = \sigma\mu q^{-1/2} - q^{1/2},$$

where $\|\sigma\mu q^{-1/2} - q^{1/2}\|^2 = \operatorname{tr}[(\sigma\mu q^{-1/2} - q^{1/2})^2] = \theta$. On the other hand,

$$\|H_d\Delta d + H_p\Delta p\|_F^2 - 2\langle H_d\Delta d, H_p\Delta p\rangle = \|H_d\Delta d\|_F^2 + \|H_p\Delta p\|_F^2. \qquad \square$$

Combining Lemmas 4.3 and 4.4, we get the bound

$$\|\Delta p\|_F\|\Delta d\|_F \leq \frac{1}{2}\|H_p^{-1}\|_F\|H_d^{-1}\|_F(\theta - 2\langle H_p\Delta p, H_d\Delta d\rangle).$$

Now, if $\langle H_p\Delta p, H_d\Delta d\rangle \geq 0$, we may drop it. Otherwise, we have for $\beta = 1$ that

$$-\langle H_p\Delta p, H_d\Delta d\rangle \leq \beta\|H_p\Delta p\|_F\|H_d\Delta d\|_F \leq \frac{\beta}{2}(\|H_p\Delta p\|_F^2 + \|H_d\Delta d\|_F^2).$$

If we can actually take $\beta < 1$, we get a geometrical series converging to the limit $(\|H_p^{-1}\|_F\|H_d^{-1}\|_F/2)\theta/(1 - \beta)$. On the other hand, if $\beta = 1$ is the only option,

[1] Each step obtains a proportional decrease of at least $1 - \hat{\delta}$ in μ, so one obtains the condition $\tau \leq (1 - \hat{\delta})^k$. Now apply the approximation $-\log(1 - \hat{\delta}) \geq \hat{\delta}$.

we have $-\langle H_p\Delta p, H_d\Delta d\rangle = \|H_p\Delta p\|_F\|H_d\Delta d\|_F$, which says that $H_0L(d)\Delta p + \tau H_0L(p)\Delta d = 0$ for some $\tau \geq 0$. That is, $L(d)\Delta p + \tau L(p)\Delta d = 0$, which means (4.2)–(4.4) must be singular. Consequently, if $\beta \nearrow 1$, (p, d) must be approaching a singularity of the system. Sufficiently far from a singularity, we thus get the following bounds.

Lemma 4.5. *Suppose that*

$$-\langle H_p\Delta p, H_d\Delta d\rangle \leq \beta\|H_p\Delta p\|_F\|H_d\Delta d\|_F$$

for $\beta < 1$. Then

$$\kappa \leq (1/\gamma + 1/r)\theta' \quad \text{and} \quad \check{\kappa} \leq (1/r)\theta'$$

for

$$\theta' \triangleq \frac{\|H_p^{-1}\|_F\|H_d^{-1}\|_F}{2(1-\beta)\mu}\theta.$$

Consequently

$$\delta^{-1} \leq 2\max\Big\{\frac{1/\gamma + 1/r}{\sigma(1-\sigma)}\theta', \frac{2/r}{(1-\sigma)^2}\theta', \frac{1}{1-\sigma}\Big\},$$

where $r = 1$ gives an upper bound for the max-term.

Proof. Note that we have both $\|\Delta\|_F \leq \|\Delta d\|_F\|\Delta p\|_F$, as remarked in Section 3.2.2, as well as $-\operatorname{tr}\Delta \leq \|\Delta d\|_F\|\Delta p\|_F$. Thus $\kappa \leq (1 + \gamma/r)\|\Delta d\|_F\|\Delta p\|_F/(\gamma\mu)$ and $\check{\kappa} \leq \|\Delta d\|_F\|\Delta p\|_F/(r\mu)$. Approximating as discussed above, and noting that $(1 + \gamma/r)/\gamma = 1/\gamma + 1/r$, yields the claimed bounds for κ and $\check{\kappa}$. Now apply these bounds in $\bar{\alpha}^{-1} = \kappa/\sigma$ ($\kappa \geq \sigma$) and $\check{\alpha}^{-1} = 2\check{\kappa}/(1-\sigma)$, and insert the results into (4.8), i.e., $\delta^{-1} \leq 2\hat{\alpha}^{-1}/(1-\sigma)$, to yield the first two terms of the maximum expression. The last term is obtained by bounding $\hat{\alpha} \leq \bar{\alpha} \leq 1$. □

The following result ensures that θ/μ stays bounded in the neighbourhoods $\mathcal{C}_\bullet$ under consideration.

Lemma 4.6. *Suppose $\|P_e^{\perp}w\|_\bullet \leq \gamma\mu(w)$ for $\gamma \in (0, 1)$, $w \in \mathcal{J}$. Then, for $\sigma > 0$,*

$$\theta(w, \sigma) \leq \Big(\frac{\gamma^2 + (1-\sigma)^2 r}{1-\gamma}\Big)\mu(w) \quad \text{when} \quad \bullet = F, \text{ and} \tag{4.9}$$

$$\theta(w, \sigma) \leq \Big(1 - 2\sigma + \frac{\sigma^2}{1-\gamma}\Big)\mu(w)r \quad \text{when} \quad \bullet = 2, -\infty. \tag{4.10}$$

Proof. See the proof of [Schmieta and Alizadeh, 2003, Lemma 35], that actually only depends on the properties of w, not of s and x (p and d). □

It remains to consider H_p and H_d.

Lemma 4.7. *Suppose $p, d, q \in \operatorname{int}\mathcal{K}$. Then,*

(i) The operators $L(q)^{-1/2}$ and $L(q^{-1/2})$ satisfy the terms of Lemma 4.4 for H_0.

(ii) When p and d operator-commute, we may take $H_0 = L(d)^{-1/2}L(p)^{-1/2}$, and get $\|H_p^{-1}\|_F\|H_d^{-1}\|_F \leq \sqrt{\mathrm{cond}(H)}$ for $H \triangleq L(d)^{-1}L(p)$.

Proof. (i) Clearly the operators are invertible. Furthermore, $L(q)^{-1/2} = L(q^{-1/2})$ on the space spanned by the eigenvectors of q. Therefore, for both alternatives, $H_0 q = q^{1/2}$ and $H_0 e = q^{-1/2}$.

(ii) Since $p, d \in \operatorname{int}\mathcal{K}$ operator-commute, H_0 is symmetric and they share a Jordan frame, wherefore $q^t = p^t \circ d^t$. Thus $H_0 q = q^{1/2}$ and $H_0 e = q^{-1/2}$. Also by operator-commutativity $H_d = H_0 L(p) = H^{1/2}$ and $H_p = H_0 L(d) = H^{-1/2}$, so that $\|H_p^{-1}\|_F\|H_d^{-1}\|_F = (\|H\|_F\|H^{-1}\|_F)^{1/2} = \sqrt{\mathrm{cond}(H)}$. □

The results of this section are summarised in the following algorithm and theorem, recalling that we may scale our representation of f_v. For $\bullet = F$, better $\sqrt{r}$ complexities could actually be obtained by limiting σ, as shown in Schmieta and Alizadeh [2003].

Algorithm 4.1 (Interior point method for DC problems on symmetric cones)**.**

1. Choose target accuracy $\underline{\mu} > 0$, parameters $\gamma, \sigma \in (0,1)$, and an initial iterate $(p,d) \in \mathcal{C}_\bullet(\gamma) \cap G_{r\mu}^{-1}(y,0)$ for some $\bullet \in \{F, 2, -\infty\}$ and $y \in \mathbb{R}^m$.
2. Choose a scaling Q_v such that $(\widetilde{p}, \underline{\widetilde{d}}) \in \mathcal{C}_\bullet^*(\gamma)$, and a H_0 satisfying the constraints of Lemma 4.4 with respect to $(\widetilde{p}, \underline{\widetilde{d}})$.
3. Solve $(\Delta\widetilde{p}, \Delta\underline{\widetilde{d}})$ from (4.2)–(4.4) if possible. Otherwise stop with failure.
4. Update $(p,d) \triangleq (Q_v^{-1}\widetilde{p}(\hat{\alpha}), Q_v\underline{\widetilde{d}}(\hat{\alpha}))$ as the new iterate.
5. If $\mu \leq \underline{\mu}$, stop. Otherwise continue from Step 2

Theorem 4.1. *Suppose that Step 3 of Algorithm 4.1 always succeeds, and there exists at each iteration an H_0 satisfying the conditions of Lemma 4.4 with respect to $(\widetilde{p}, \underline{\widetilde{d}})$. Suppose furthermore that $\|H_p^{-1}\|_F\|H_d^{-1}\|_F/(1-\beta)$ can be bounded from above by a constant $M < \infty$. Denote by $\overline{\mu}$ the initial (maximal) μ and let $\tau \triangleq \underline{\mu}/\overline{\mu}$. Then $O(Mr\log\tau^{-1})$ iterations are sufficient for $\mu \leq \underline{\mu}$.*

Proof. Note that since $\mathcal{C}_\bullet^*(\gamma) \subset \mathcal{C}_\bullet(\gamma)$, and the latter is scaling invariant, after reverse scaling still $(Q_v^{-1}\widetilde{p}(\alpha), Q_v\underline{\widetilde{d}}(\alpha)) \in \mathcal{C}_\bullet(\gamma)$. Therefore Step 4 and the method are well-defined.

Other dependencies on r in the bound for δ^{-1} from Lemma 4.5 can be approximated away, except the linear one in (4.9) or (4.10). Thus $\delta^{-1} = O(Mr)$, and the claim follows from the discussion following Lemma 4.2. □

4.2.5 Operator-commutative scalings

Suppose we choose the scaling such that $\widetilde{p} = Q_v p$ and $\underline{\widetilde{d}} = Q_v^{-1}d$ operator-commute. As discussed in Section 4.2.3, then $(p,d) \in \mathcal{C}_\bullet(\gamma)$ implies $(\widetilde{p}, \underline{\widetilde{d}}) \in \mathcal{C}_\bullet^*(\gamma)$, taking care of that assumption in Theorem 4.1. Lemma 4.7 then says that it remains to bound $\mathrm{cond}(H)$ (and stay away from a singularity).

In the Nesterov-Todd method, the scaling element is chosen to be v for the unique element for which $Q_{v^2}p = d$, expressible as $v = (Q_{p^{1/2}}(Q_{p^{1/2}}d)^{-1/2})^{-1/2}$ according to Schmieta and Alizadeh [2003]. Then $\widetilde{p} = \underset{\sim}{d}$ operator-commute, and $L(\underset{\sim}{d})^{-1}L(\widetilde{p}) = I$, so that consequently $\mathrm{cond}(H) = 1$. In the so-called "xs" method, $v = d^{1/2}$, so that $\underset{\sim}{d} = e$, wherefore we have operator-commutativity, and get $\mathrm{cond}(H) \leq 2/(1-\gamma)$ for $\bullet = 2, F$, and $\mathrm{cond}(H) \leq r/(1-\gamma)$ for $\bullet = -\infty$. In the "sx" method $v = p^{-1/2}$, with similar results. More generally, the so-called power class of scalings (or search directions) considered by Muramatsu [2002], yields bounded $\mathrm{cond}(H)$.

Of course, the question remains: what is the effect of scaling on the closeness of the system (4.2)–(4.4) to a singularity? By the discussion following Lemma 3.6, this is at least somewhat unaffected close to a central selection. Also, when A and B_- have the special forms of Assumption 3.2 and Lemma 3.9, Lemmas 3.14 and 3.15 show that (strict) complementarity and non-degeneracy are unaffected by scaling. Therefore, Lemma 3.6 shows that any scaled representation is non-singular in some neighbourhood of a point that satisfies additional scaling-independent (for $q = 0$) second-order assumptions.

4.3 Globalisation: A filter method

4.3.1 The idea

The idea of the filter method was first introduced for constrained optimisation by Fletcher and Leyffer [2002] in a sequential quadratic programming (SQP) framework, with convergence proven in Fletcher et al. [2002], for the case considered. Other works in filter algorithms that seem most related to our work include those of Ulbrich et al. [2004] and Wächter and Biegler [2005], where interior point approaches are considered.

The filter is basically a multi-dimensional generalisation of a monotonically decreasing sequence bounded from below, where the decrease at each step is sufficient by some criterion. Each point inserted in the filter defines a cone of other points it dominates. Points belonging in an envelope of such a cone are not allowed in future iterations. A filter method is therefore multi-objective optimisation applied to single-objective problems, where typically the additional objectives are related to the constraints of the problem.

In practical methods in the literature so far, there are only two objectives, and each of them is improved separately. One of them, typically the original objective function value, is assigned to be the primary objective, and decrease in it is sought while allowed by the filter, and some additional sufficient decrease conditions are met. New points are inserted in the filter at appropriate places, to force convergence in the future. When this primary phase of the algorithm runs into trouble, a *restoration phase* is entered, with the purpose of improving the second objective and restoring feasibility and acceptability to the filter. Often this

restoration method is taken to be a black box.

The restoration method in Ulbrich et al. [2004], however, is closely related to the primary method, and merely advances slightly differently. Indeed, although rather general (C^2) constrained nonlinear programming is considered therein, the resulting analysis bears many parallels to the work in Section 4.2, and more generally the work on linear programming on symmetric cones. Their two elements of the filter actually include the values $\mu(p \circ d)$ and $\|P_e^{\perp}(p \circ d)\|$ (in the non-negative orthant of $\mathbb{R}^m$, instead of general symmetric cones), plus additional terms related to dissatisfaction of linear constraints. However, to prove convergence for the filter method, it is assumed that the equivalent of the system (4.2)–(4.4) is suitably far from a singularity. But with such assumptions, the methods of Section 4.2 do already converge, "fast". It is our intent to use the idea of the filter method to circumvent that assumption. We will use a filter and a restoration method to restore feasibility, when the main interior point method runs into trouble. To do this, we apply the results of Section 4.4 to follow, as a consequence of which our restoration phase algorithm will also be closely related to the primary phase algorithm.

4.3.2 The method

We take the filter $\mathcal{F}$ to be a set of pairs $(g^{\mathcal{F}}, h^{\mathcal{F}}) \in \mathbb{R} \times [0, \infty)$. Then, another point (g, h) is considered *acceptable to the filter* if for prescribed values of $\delta_{\mathcal{F}} \in (0, 1)$ and $\theta_{\mathcal{F}} > 0$,

$$\text{for all } (g^{\mathcal{F}}, h^{\mathcal{F}}) \in \mathcal{F} \quad \text{either} \quad g \leq g^{\mathcal{F}} - \theta_{\mathcal{F}} h^{\mathcal{F}} \quad \text{or} \quad h \leq (1 - \delta_{\mathcal{F}}) h^{\mathcal{F}}.$$

By *augmenting the filter* with (g, h) we mean replacing it with

$$\{(g, h)\} \cup \{(g^{\mathcal{F}}, h^{\mathcal{F}}) \in \mathcal{F} \mid g^{\mathcal{F}} < g \text{ or } h^{\mathcal{F}} < h\}.$$

The first part of the following lemma is standard:

Lemma 4.8. *Suppose points added to the filter satisfy $g \in [\underline{g}, \overline{g}] \subset (-\infty, \infty)$ and $h \geq \underline{h} > 0$. Then the filter may be augmented only finitely many times with acceptable points (g, h). If, furthermore, $h \leq \overline{h}$, then the filter may be augmented at most $[(\overline{g} - \underline{g})/(\theta_{\mathcal{F}} \underline{h}) + 1][\delta_{\mathcal{F}}^{-1} \log \tau^{-1} + 1]$ times for $\tau \triangleq \underline{h}/\overline{h}$. In particular, if $\overline{g} - \underline{g} = O(\overline{h})$, then we have the bound $O(\tau^{-1} \log \tau^{-1})$ for the number of augmentations.*

Proof. Consider the square $A \triangleq [\underline{h}, \overline{h}] \times [\underline{g}, \overline{g}]$. It is covered by the rectangles $(\overline{h}(1 - \delta_{\mathcal{F}})^n [1 - \delta_{\mathcal{F}}, 1]) \times (\overline{g} - \theta_{\mathcal{F}} \underline{h}[k, k + 1])$, where $n = 0, 1, \ldots, N - 1$, and $k = 0, 1, \ldots, K - 1$. At most one point acceptable to the filter may lie in each rectangle, so the number of rectangles KN gives an upper bound on the number of acceptable points that may be inserted in the filter. Solving $\underline{g} > \overline{g} - \theta_{\mathcal{F}} \underline{h} K$, we get $K > (\overline{g} - \underline{g})/(\theta_{\mathcal{F}} \underline{h})$. Solving for N from $\underline{h} > (1 - \delta_{\mathcal{F}})^N \overline{h}$, we get the *sufficient* condition $N > \log(\overline{h}/\underline{h}) \delta_{\mathcal{F}}^{-1}$ (by application of $-\log(1 - \delta_{\mathcal{F}}) \geq \delta_{\mathcal{F}}^2/2 + \delta_{\mathcal{F}} \geq \delta_{\mathcal{F}}$). This gives the desired bound in the case $h \leq \overline{h}$.

Suppose then that there is an infinite sequence $(h_{[k]}, g_{[k]}) \in \mathcal{F}$, $k = 1, 2, \ldots$, with $h_{[k+1]} \geq h_{[k]}$. Then $g_{[k+1]} \leq g_{[k]} - \theta_{\mathcal{F}} h_{[k]} \leq g_{[k]} - \theta_{\mathcal{F}} \underline{h}$, so that $g_{[k+1]} \leq g_{[1]} - k\theta_{\mathcal{F}} \underline{h}$, and for large enough k, $g_{[k+1]} < \underline{g}$, which is a contradiction. Therefore there exists some finite $\overline{h} \geq h$, and only finitely many entries may be added in the filter. □

In our present situation, we take $g = f_\nu(y)$ as the quality of the solution in terms of objective function value, and $h = \epsilon = r\mu$ as the quality of the solution in terms of $0 \in \partial_\epsilon^{\mathrm{DC}} f_\nu(y)$, as in Algorithm 4.1. Therefore, in contrast to the situation in constrained optimisation, either filter element becoming sufficiently small provides an approximate solution of prescribed quality. Unless the restoration method fails (which our restoration method of choice will not do), it always generates either a point acceptable to the filter, or a solution of such prescribed quality, by reducing the value of f_ν or ϵ sufficiently. Therefore Lemma 4.8 alone proves convergence of the filter method in case of non-failure, if we augment the filter with acceptable points between restoration steps. Hence, the primary design goal of the filter method is to obtain greater (in practise) convergence speeds than the pure restoration method.

A crude filter method would therefore simply augment the filter and enter the restoration phase, whenever the main interior point method does not provide sufficient decrease in ϵ (or sufficiently long step), or the candidate iterate is unacceptable to the filter. Limited practical experience, however, suggests that an approach familiar from other filter methods in the literature works better. We next represent such a method. The idea is to choose a shorter step size than allowed by the pure interior point method, if f_ν is sufficiently descending in the search direction. Also, if a linear model of the function does not predict decrease, we augment the filter for future reference.

In the rest of this section, we assume that both f and ν are of the form (3.7).

Suppose $y, \Delta y \in \mathbb{R}^m$ are given, and $0 \in \partial_\epsilon^{\mathrm{DC}} f(y)$. For arbitrary $z \in \partial^{\mathrm{DC}} f_\nu(y)$, we define the linear model of f_ν,

$$l(\alpha) \triangleq f_\nu(y) + \alpha \langle z, \Delta y \rangle.$$

We say that the model decreases sufficiently, if for prescribed $\kappa > 0$,

$$l(0) - l(\alpha) \geq \kappa\epsilon, \tag{4.11}$$

and that f_ν itself decreases sufficiently with respect to the model, if for given $\eta > 0$,

$$\frac{f_\nu(y) - f_\nu(y(\alpha))}{l(0) - l(\alpha)} \geq \eta. \tag{4.12}$$

Here we denote $y(\alpha) \triangleq y + \alpha\Delta y$ akin to (4.5). We also introduce the notation $\epsilon(\alpha) \triangleq \langle p(\alpha), d(\alpha) \rangle = r\mu(\alpha)$, where $\mu(\alpha)$ is given by (4.6).

With these definitions, the filter method is as follows.

Algorithm 4.2 (Filter method for DC problems on symmetric cones)**.**

1. Choose target accuracy $\underline{\epsilon} > 0$, parameters $\delta, \delta_{\mathcal{F}} \in (0,1)$, $\theta_{\mathcal{F}} > 0$, $\eta \in (0,1)$, and $\kappa > 0$, as well as the filter $\mathcal{F}$ and its initial contents.

2. Initialise the interior point method per instructions of Algorithm 4.1 for the data of f_v, yielding (p, d, y, ϵ) with $(p, d) \in G_\epsilon^{-1}(y, 0) \cap \mathcal{C}_\bullet(\gamma)$.

3. If $\epsilon \leq \underline{\epsilon}$, stop, for we have a solution.

4. Calculate the direction $(\Delta p, \Delta d, \Delta y)$ by solving, as in Algorithm 4.1, a scaled version of (4.2)–(4.4). Set $\alpha \triangleq \hat{\alpha}$ with the latter as in Lemma 4.2.

5. If Step 4 failed, or $\epsilon(\alpha)/\epsilon > 1 - \delta$, augment $\mathcal{F}$ with $(f_v(y), \epsilon)$, and enter the restoration phase that either

   (a) Produces a new iterate (p, d, y, ϵ) with $(p, d) \in G_\epsilon^{-1}(y, 0) \cap \mathcal{C}_\bullet(\gamma)$ and $(f_v(y), \epsilon)$ acceptable to the filter. In this case we continue from Step 4.

   (b) Detects an $\underline{\epsilon}$-semi-critical point (or fails), in which case we stop.

6. If $(f_v(y(\alpha)), \epsilon(\alpha))$ is acceptable to $\mathcal{F}$, and either (4.11) fails or (4.12) holds, go to Step 8.

7. Set $\alpha \triangleq \alpha/2$, and go to Step 5.

8. If (4.11) fails, augment $\mathcal{F}$ with $(f_v(y), \epsilon)$.

9. Update $(p, d, y, \epsilon) \triangleq (p(\alpha), d(\alpha), y(\alpha), \epsilon(\alpha))$, and continue from Step 3.

Theorem 4.2. *Suppose the filter $\mathcal{F}$ is initialised to include $\{(\overline{g}, 0)\}$ for some $\overline{g} > \min f_v$ (and that the initial iterate is acceptable to $\mathcal{F}$). Then Algorithm 4.2 converges in a finite number of iterations to an $\underline{\epsilon}$-semi-critical point (if the restoration method does not fail). If, furthermore, always $\epsilon \leq \overline{\epsilon}$ for some $\overline{\epsilon} > \underline{\epsilon}$ such that $\overline{\epsilon} > \overline{g} - \min f_v$, and the restoration method is taken as an oracle, then the number of iterations is $O(\tau^{-1}(\log \tau^{-1})^2)$ for $\tau \triangleq \underline{\epsilon}/\overline{\epsilon}$.*

Proof. Step 5 ensures $\epsilon(\alpha)/\epsilon \leq 1 - \delta$. Thus a standard argument (cf. Lemma 4.8) shows that there are at most $O(\log \tau^{-1})$ iterations of the main phase of the algorithm between each restoration phase. Since the filter is augmented before each restoration phase with a point acceptable to it, Lemma 4.8 says that the restoration method may be called only a finite number of times. Furthermore, when $\epsilon \leq \overline{\epsilon}$, Lemma 4.8 with $\underline{g} = \min f_v$ provides the bound $O(\tau^{-1} \log \tau^{-1})$ for the number of augmentations. □

Remark 4.1. Instead of directly specifying δ, we could specify $\beta \in (0,1)$, and calculate δ^{-1} according to Lemma 4.5. In this case we should include in the complexity estimate, the contribution by r, and potentially γ as well, depending on whether reinitialisation of $(p, d) \in G_\epsilon^{-1}(y, 0) \cap \mathcal{C}_\bullet(\gamma)$ in the restoration method allows free choice, or guarantees a bound.

4.4 The restoration method

4.4.1 Sequential convex programming

Consider two arbitrary finite convex functions f and ν on $\mathbb{R}^m$. Let $\underline{\epsilon} \geq 2\rho \geq 0$ be chosen. Suppose $z \in \partial_\rho \nu(y)$, $z \notin \partial_{\underline{\epsilon}-\rho} f(y)$. In other words,

$$\nu(y') - \nu(y) \geq z^T(y' - y) - \rho, \quad \text{for all } y', \tag{4.13}$$

$$f(y'') - f(y) < z^T(y'' - y) - (\underline{\epsilon} - \rho), \quad \text{for some } y''. \tag{4.14}$$

Setting $y' = y''$ and summing,

$$f_\nu(y'') - f_\nu(y) < -\underline{\epsilon} + 2\rho$$

so that y is not $\underline{\epsilon} - 2\rho$ -optimal.

Suppose then that we have $z \in \partial_{\epsilon'} f(\hat{y})$, i.e.,

$$f(y') - f(\hat{y}) \geq z^T(y' - \hat{y}) - \epsilon', \quad \text{for all } y'.$$

Setting $y' = y''$, and summing with (4.14), we have

$$f(y) - f(\hat{y}) > z^T(y - \hat{y}) - \epsilon' + (\underline{\epsilon} - \rho).$$

Setting $y' = \hat{y}$ and further summing with (4.13),

$$f_\nu(y) - f_\nu(\hat{y}) > (\underline{\epsilon} - 2\rho) - \epsilon'. \tag{4.15}$$

Thus, if $\epsilon' \leq \sigma_{\text{SCP}}(\underline{\epsilon} - 2\rho)$ for $\sigma_{\text{SCP}} \in (0, 1)$, a reduction of $(1 - \sigma_{\text{SCP}})(\underline{\epsilon} - 2\rho)$ has been achieved in the value of f_ν.

The conceptual algorithm for finding $\underline{\epsilon}$-semi-critical points of f_ν is now clear.

Algorithm 4.3 (Sequential convex programming (SCP) method)**.**

1. Choose target accuracy $\underline{\epsilon} > 0$, gradient accuracy $\rho \in [0, \underline{\epsilon}/2)$, stepwise reduction $\sigma_{\text{SCP}} \in (0, 1)$, and an initial iterate $y_{[0]}$.

2. Select a subgradient $z_{[k]} \in \partial_\rho \nu(y_{[k]})$.

3. Set $\epsilon' \triangleq \sigma_{\text{SCP}}(\underline{\epsilon} - 2\rho)$, and find $\hat{y}$ such that $z_{[k]} \in \partial_{\epsilon'} f(\hat{y})$.

4. If a reduction of $(1 - \sigma_{\text{SCP}})(\underline{\epsilon} - 2\rho)$ is not obtained in the value of f_ν, by the above analysis it must have been that $z_{[k]} \in \partial_{\underline{\epsilon}-\rho} f(y_{[k]})$, so that $0 \in \partial^{\text{DC}}_{\underline{\epsilon}} f_\nu(y_{[k]})$, and we already were at a $\underline{\epsilon}$-semi-critical point. Therefore, stop with result $y_{[k]}$.

5. Otherwise repeat from Step 2 with $y_{[k+1]} \triangleq \hat{y}$, and $k \triangleq k + 1$.

Clearly, as a constant reduction in the value of f_ν is achieved on each non-final iteration, the method is convergent if Step 3 always succeeds, and f_ν is bounded from below. For the success, we should have $\mathcal{R}(\partial\nu) \subset \mathcal{R}(\partial f)$. The stricter bound $\operatorname{cl}\mathcal{R}(\partial\nu) \subset \operatorname{int}\mathcal{R}(\partial f)$ along with bounded $\mathcal{R}(\partial\nu)$ in fact ensures that f_ν has bounded level sets and is therefore bounded from below according to Theorem 2.6.

We note that this method can be seen as an approximate variant of the DCA method of An and Tao [2005], the "simplified" version of which amounts to $\rho = \underline{\epsilon} = 0$ (while the "complete" version sets further restrictions). The method of truncated codifferential considered by Demyanov et al. [2002] also bears many parallels to SCP.

Remark 4.2. Alternatively, instead of fixing ϵ' in Step 3, we may attempt to find $\hat{y}$ and $\epsilon' > 0$ with $z \in \partial_{\epsilon'} f(\hat{y})$, such that the objective function value is reduced by $0 < \Delta_{[k]} \leq \underline{\epsilon} - 2\rho$, or (4.15) is violated (for $y = y_{[k]}$), one of which must occur for small enough $\epsilon' > 0$.

Remark 4.3. The SCP argument actually proves convergence for inexact K-means -style local convex optimisation methods; cf. Section 6.3 and the references therein. Suppose $f_\nu(y) = f(y) - \nu(y)$ for $\nu(y) \triangleq \max_{t\in T} \nu_t(y)$ for some finite index set T and convex functions ν_t, and that $f_t \triangleq f - \nu_t$ are convex. Now, suppose $f_\nu(y) = f_t(y)$, and choose in the SCP method, f_t for f, 0 for ν, $z = 0$ and $\rho = 0$. If the predicted decrease is not achieved, then the SCP argument says $0 \in \partial_{\underline{\epsilon}} f_t(y)$, that is $f(y') - f(y) \geq \nu_t(y') - \nu_t(y) - \underline{\epsilon}$ for all y'. But then for any $z \in \partial\nu_t(y) \subset \partial\nu(y)$, $f(y') - f(y) \geq z^T(y' - y) - \underline{\epsilon}$. This says $z \in \partial_{\underline{\epsilon}} f(y)$, so that y is $\underline{\epsilon}$-semi-critical for f_ν.

4.4.2 Interior point SCP

If f (but not necessarily ν) has the form (3.7), we may apply Algorithm 4.1 in Step 3 of Algorithm 4.3 to reducing $\epsilon' > 0$ in $z \in \partial_{\epsilon'} f(\hat{y})$, after finding initial values for which this holds. For, as is clear from the analysis, Algorithm 4.1 always maintains the linear constraints for any set values, and therefore works for other values besides $z = 0$. If we can initialise each iteration in a bounded manner, we have finite convergence. More precisely,

Theorem 4.3. *Suppose that for all $y_{[k]}$ and $z_{[k]} \in \partial_\rho \nu(y_{[k]})$, we can (in negligible time) initialise $(p_f, d_f) \in G^{-1}_{f,\overline{\epsilon}}(y', z_{[k]}) \cap \mathcal{C}_\bullet(\gamma)$ at some y' for fixed $\overline{\epsilon} \geq f_\nu(y_{[0]}) - \min f_\nu$, $\gamma \in (0,1)$, and $\bullet \in \{F, 2, -\infty\}$. Then, if Algorithm 4.1 is used in Step 3 with one of the operator-commutative scalings from Section 4.2.5, $O(K_{\gamma,r_f}\tau^{-1}\log\tau^{-1})$ steps of the interior point method are sufficient to reach an $\underline{\epsilon}$-semi-critical point, with $\tau \triangleq (\underline{\epsilon} - 2\rho)/\overline{\epsilon}$, and K_{γ,r_f} a polynomial of $1/(1-\gamma)$ and r_f,*

Here and in the rest of this section, $G^{-1}_{f,\epsilon}$ is G^{-1}_ϵ as defined in Corollary 3.2 for the data of f, while without the specifier, the data of all of f_ν is implied, as before. $\mathcal{C}_\bullet(\gamma)$ is a subset of one of $\mathcal{K}$, $\mathcal{K}_f$, or $\mathcal{K}_\nu$, depending on the context.

The factor K_{γ,r_f} replaces Mr_f and omitted terms from Theorem 4.1, where the dependence on γ was de-emphasised, being something that can be chosen arbitrarily small by suitable initialisation. Here, however, z limits the quality of the initialisation – which cannot be done if $z \notin \mathcal{R}(\partial f)$.

Proof. The term $1/(1-\gamma)$ is the dominant one involving γ as $\gamma \nearrow 1$ in the bounds of Lemma 4.6 and the bounds for $\operatorname{cond}(H)$ in Section 4.2.5. Therefore, similarly to the proof of Theorem 4.1, we find from Lemmas 4.5 and 4.2 that to find an $\underline{\epsilon} - 2\rho$ critical point, each invocation of Step 3 requires $O(K_{\gamma,r_f} \log \tau^{-1})$ steps of the interior point method, where K_{γ,r_f} is as claimed.

Since each non-terminal step of the SCP algorithm achieves a reduction of at least $(1-\sigma_{\mathrm{SCP}})(\underline{\epsilon} - 2\rho)$ in the value of f_ν, and $\Delta_0 \triangleq f_\nu(y_{[0]}) - \min f_\nu \leq \overline{\epsilon}$, we get that $n \geq \Delta_0/((1-\sigma_{\mathrm{SCP}})(\underline{\epsilon} - 2\rho)) = O(\tau^{-1})$ iterations of the SCP method are sufficient. This results in the claimed total number of iterations of the interior point method. □

Next we study when the initialisation required above can be performed, and with what quality. We begin with a few basic results needed towards that end.

Lemma 4.9. *Suppose f_ν is bounded from below, $\rho \geq 0$, $\Delta_0 \geq f_\nu(y) - \min f_\nu$, and $z \in \partial_\rho \nu(y)$. Then $z \in \partial_{\Delta_0+\rho} f(y)$.*

Proof. By assumption $\rho \geq \nu(y) - \nu(y') + z^T(y' - y)$ and $\Delta_0 \geq f(y) - \nu(y) - f(y') + \nu(y')$ for all y'. By combining these inequalities, we get the claim. □

Lemma 4.9 and (3.11) thus show the existence of some $(p_f, d_f) \in G^{-1}_{f,\Delta_0+\rho}(y_{[k]}, z_{[k]})$. The objective is then to improve $(p_f, d_f) \in \mathcal{C}_\bullet(\gamma)$ for fixed $\gamma \in (0,1)$ without $\overline{\epsilon} \geq \Delta_0 + \rho$ increasing unboundedly.

To provide such results, we need to show that $\|\cdot\|_{-\infty}$ actually satisfies the triangle inequality (although it is not a norm).

Lemma 4.10. *Suppose $x, y \in \mathcal{J}$. Then $\|x+y\|_{-\infty} \leq \|x\|_{-\infty} + \|y\|_{-\infty}$.*

Proof. As defined in Section 4.2.3, $\|z\|_{-\infty} = -\min_i \zeta_i(z)$, so it suffices to show $\min_i \zeta_i(z) \geq \min_i \zeta_i(x) + \min_i \zeta_i(y)$ for $z = x + y$. Let $x = \sum_{i=1}^r \zeta_i(x) x_i$, $y = \sum_{i=1}^r \zeta_i(y) y_i$, and $z = \sum_{i=1}^r \zeta_i(z) z_i$ be the decompositions of $x, y, z \in \mathcal{K}$ into sums of primitive idempotents. Applying $\sum_j x_j = e$, $\operatorname{tr} z_i = 1$, and the fact that since primitive idempotents are in $\mathcal{K}$, their inner product is non-negative, we have

$$\begin{aligned}\zeta_i(z) = \langle z_i, z\rangle = \langle z_i, x\rangle + \langle z_i, y\rangle = \sum_j (\zeta_j(x)\langle z_i, x_j\rangle + \zeta_j(y)\langle z_i, y_j\rangle) \\ \geq \min_k \zeta_k(x)\langle z_i, \sum_j x_j\rangle + \min_k \zeta_k(y)\langle z_i, \sum_j y_j\rangle = \min_j \zeta_j(x) + \min_j \zeta_j(y). \quad \square\end{aligned}$$

Assumption 4.1. We assume that $A(p_1, \ldots, p_n) = (\langle a'_1, p_1\rangle, \ldots, \langle a'_n, p_n\rangle)$ as in Assumption 3.2, along with $(\mathcal{R}(A^*) \cap \operatorname{int} \mathcal{K})^{-1} \subset \mathcal{N}(B) \cap \mathcal{N}(\langle c, \cdot\rangle)$.

Remark 4.4. Inversion in the latter condition can always be made unnecessary by scaling with $v = (a'_1, \ldots, a'_n)$ to yield $\underset{\sim}{A}\widetilde{p} = (\langle e, \widetilde{p}_1\rangle, \ldots, \langle e, \widetilde{p}_n\rangle)$; cf. Section 3.4.6.

Example 4.1. This assumption is satisfied by combinations of Euclidean norms (cf. Example 3.3), where of $p_i = (p_i^0, \bar{p}_i) \in \mathcal{E}_{m+1}$, A depends only on p_i^0, and B and $\langle c, \cdot \rangle$ on $\bar{p}_i$.

When the assumption holds, we set $a \triangleq (\phi_1 a_1', \ldots, \phi_n a_n')$, where $\phi_i \in \mathbb{R}$ is chosen so that $Aa^{-1} = b$, i.e., $r_i = \phi_i b_i$. Then $a \in \mathcal{R}(A^*) \cap \operatorname{int} \mathcal{K}$, so that $\langle a^{-1}, B^* y \rangle = \langle a^{-1}, c \rangle = 0$. Also, $\mu(a \circ p) = 1$ for any $p \in V = \{p \in \mathcal{K} \mid Ap = b\}$, because $\langle a, p \rangle = \sum_i \phi_i \langle a_i', p_{f,i} \rangle = \sum_i \phi_i b_i = \sum_i r_i$.

Lemma 4.11. *Suppose Assumption 4.1 holds, and that* $(p', d') \in G_{\epsilon'}^{-1}(y, z) \cap \mathcal{C}_{-\infty}(\gamma')$. *Then, for* $0 < \psi < \gamma \leq \gamma'$, *there exist* $(p, d) \in G_{\epsilon}^{-1}(y, \psi z) \cap \mathcal{C}_{-\infty}(\gamma)$ *with*

$$\epsilon \triangleq \psi \frac{\gamma' - \psi}{\gamma - \psi} \langle p', d' \rangle + \frac{(1-\psi)^2}{\gamma - \psi} \langle a^{-1}, d' \rangle \leq \frac{1 + (\gamma' - 2)\psi}{\gamma - \psi} \epsilon' + \frac{(1-\psi)^2}{\gamma - \psi} v(y), \tag{4.16}$$

where $v(y) \triangleq \sup_{p \in V} \langle p, B^* y + c \rangle$. *In particular, with* $\gamma' = 1$ *and* $\gamma = (1 + \psi)/2$, *we get* $\epsilon = 2\langle p, d' \rangle$ *and* $1/(1 - \gamma) = O(1/(1 - \psi))$.

By the definition of f and v, when the lemma is applied to the data of f alone, $v = f$, and when it is applied to f_v, $v = f + v$.

Proof. Letting $p \triangleq \psi p' + (1 - \psi) a^{-1}$, we have $Bp = \psi z$, and by convexity $p \in V$. Furthermore, $Q_a^{1/2} p = Q_a^{1/2}(\psi p') + (1 - \psi) e$, so that $P_e^{\perp} Q_a^{1/2} p = \psi P_e^{\perp} Q_a^{1/2} p'$. Since $Q_a^{1/2} p' \in \mathcal{K}$, we have $\min_i \zeta_i(Q_a^{1/2} p') \geq 0$, and then

$$\|P_e^{\perp} Q_a^{1/2} p\|_{-\infty} = \psi \|P_e^{\perp} Q_a^{1/2} p'\|_{-\infty} \leq \psi \mu(Q_a^{1/2} p') = \psi = \psi \mu(a \circ p). \tag{4.17}$$

Now, let $d \triangleq d' + \lambda a$, for yet unspecified $\lambda \geq 0$. Clearly $d \in \mathcal{K}$. Now $Q_p^{1/2} d = \lambda Q_p^{1/2} a + Q_p^{1/2} d'$, and both of the components are in $\mathcal{K}$. Therefore, we may apply Lemma 4.10 and get by the symmetricity $\|P_e^{\perp} Q_p^{1/2} a\|_{-\infty} = \|P_e^{\perp} Q_a^{1/2} p\|_{-\infty}$ [Schmieta and Alizadeh, 2003, Proposition 21] that

$$\begin{aligned} \|P_e^{\perp} Q_p^{1/2} d\|_{-\infty} &\leq \lambda \|P_e^{\perp} Q_p^{1/2} a\|_{-\infty} + \|P_e^{\perp} Q_p^{1/2} d'\|_{-\infty} \\ &\leq \lambda \|P_e^{\perp} Q_p^{1/2} a\|_{-\infty} + \psi \|P_e^{\perp} Q_{d'}^{1/2} p'\|_{-\infty} + (1 - \psi) \|P_e^{\perp} Q_{d'}^{1/2} a^{-1}\|_{-\infty} \\ &\leq \lambda \psi + \psi \gamma' \mu(p' \circ d') + (1 - \psi) \mu(a^{-1} \circ d'). \end{aligned}$$

Since

$$\mu(p \circ d) = \lambda \mu(p \circ a) + \mu(p \circ d') = \lambda + \psi \mu(p' \circ d') + (1 - \psi) \mu(a^{-1} \circ d'), \tag{4.18}$$

we therefore have $\|P_e^{\perp} Q_p^{1/2} d\|_{-\infty} \leq \gamma \mu(d \circ p)$, if

$$\psi(\gamma' - \gamma) \mu(p' \circ d') + (1 - \psi)(1 - \gamma) \mu(a^{-1} \circ d') \leq (\gamma - \psi) \lambda.$$

Setting this to equality and inserting the resulting λ in (4.18), gives the first half of (4.16) (as $\epsilon = r\mu(p \circ d)$).

For the second half of (4.16), observe that $\langle a^{-1}, d' \rangle = \langle p', d' \rangle + \langle a^{-1} - p', d' \rangle = \langle p', d' \rangle + \langle p' - a^{-1}, B^* y + c \rangle = \langle p', d' \rangle + \langle p', B^* y + c \rangle \leq \epsilon + v(y)$ by Assumption 4.1.

Finally, setting $\gamma' = 1$ and $\gamma = (1+\psi)/2$, we have $\gamma - \psi = (1-\psi)/2$, and therefore $\epsilon = 2(\psi\langle p', d'\rangle + (1-\psi)\langle a^{-1}, d'\rangle)$. By the definition of p, this proves the claim for that case. □

Lemma 4.12. *Suppose Assumption 4.1 holds for f, and that $\mathcal{R}(\partial\nu) \subset \psi\mathcal{R}(\partial f)$ for some $\psi \in (0,1)$. Then there exist $(p_f, d_f) \in G^{-1}_{f,\bar{\epsilon}}(y,z) \cap \mathcal{C}_{-\infty}(\gamma)$, $\gamma \in (0,1)$, with $1/(1-\gamma) = O(1/(1-\psi))$, in the following cases:*

(i) Varying y with $f_\nu(y) - \min f_\nu \le \Delta_0$ and $z \in \partial_\rho \nu(y)$, in which case $\bar{\epsilon} = O(\Delta_0 + \rho + \|V_f\|_F \|c_f\|_F)$.

(ii) Fixed y with $z \in \mathcal{R}(\partial\nu)$, in which case $\bar{\epsilon} = O(f(y)) = O(\|V_f\|_F \|B_f^* y + c_f\|_F)$.

As usual, the set norm is defined as $\|V_f\|_F \triangleq \max_{p\in V_f} \|p\|_F$.

Proof. Note that $\operatorname{cl}\mathcal{R}(\partial f) = \operatorname{cl}\bigcup_{y\in\mathbb{R}^m, \epsilon\ge 0} \partial_\epsilon f(y) = B_f V_f$, also from the representation of (3.11). Therefore, for $z \in \mathcal{R}(\partial\nu)$, there exists $p' \in V_f$ such that $B_f p' = z/\psi$. An application of Lemma 4.11 to (p', d') and z/ψ with $\gamma' = 1$ and $\gamma = (1+\psi)/2$ then provides $(p_f, d_f) \triangleq (p, d)$ and the requested bounds as follows:

(i) Let $(p'', d') \in G^{-1}_{f,\Delta_0+\rho}(y,z)$ as shown to exist by Lemma 4.9 and the representation of (3.11). Now, for the p provided by Lemma 4.11 we approximate $\langle p, d'\rangle = \langle p'', d'\rangle + \langle p - p'', d'\rangle = \langle p'', d'\rangle + \langle p'' - p, c_f\rangle \le \Delta_0 + \rho + 2\|V_f\|_F\|c_f\|_F$, where in the second equality we have used $B_f p = B_f p'' = z$ and $A_f p = A_f p'' = b$.

(ii) Choose $(p'', d') \in G^{-1}_{f,0}(y, z')$ for some $z' \in \partial f(y)$. Then, as in case i), $\langle p, d'\rangle = \langle p'', d'\rangle + \langle p - p'', d'\rangle = \langle p'' - p, B_f^* y + c_f\rangle$, and we readily get the claim by the definition of f. □

According to Lemma 4.12, there then is a solution to our initialisation problem under rather reasonable assumptions; cf. the level-boundedness results of Section 2.5. But when can we actually find p such that $B_f p = z/\psi$ in V_f? Since $\operatorname{tr}(Q_a^{1/2} p)$ is constant, by the proof of Lemma 4.11, $\|Q_a^{1/2} p\|_{-\infty}$ can be made small enough to imply that $p \in \mathcal{K}$. Therefore, after scaling by a to work on $\widetilde{p} \triangleq Q_a^{1/2} p$, and relaxing the norm to $\bullet \in \{F, 2, -\infty\}$, this problem may be cast as $\min_{\widetilde{p}} \|\widetilde{p}\|_\bullet$ subject to $W\widetilde{p} = x_\psi$ and $\widetilde{p} \in \mathcal{K}$, where $W\widetilde{p} \triangleq (A_f Q_a^{-1/2}\widetilde{p}, B_f Q_a^{-1/2}\widetilde{p})$ and $x_\psi \triangleq (b, z/\psi)$.

If $\bullet = -\infty$, there exists an interior solution for non-minimal ψ. The problem then becomes $\min_{\widetilde{p}}(-\min_j \zeta_j(\widetilde{p})) = \min_{\widetilde{p}} \max_j(-\zeta_j(\widetilde{p}))$. If f has the product presentation of Assumption 4.1, and each of the cones $\mathcal{K}_i$ are second-order cones, the smallest eigenvalue in each cone is $\widetilde{p}_i^0 - \|\widetilde{\widetilde{p}}_i\|$. But $\widetilde{p}_i^0$ is fixed because $b = A_f Q_a^{-1/2}\widetilde{p} = (\phi_1^{-1}\langle e, \widetilde{p}_1\rangle, \ldots, \phi_n^{-1}\langle e, \widetilde{p}_2\rangle) = (b_1\widetilde{p}_1^0, \ldots, b_n\widetilde{p}_n^0)$. Therefore the problem becomes $\min_{\widetilde{p}} \max_i \|\widetilde{\widetilde{p}}_i\|$ subject to the linear constraints.

If we set $\bullet = F$, we have $\widetilde{p} = W^\dagger x_\psi$ for the Moore-Penrose pseudo-inverse $W^\dagger = W^*(WW^*)^{-1}$ (as by assumption $\mathcal{N}(W^*) = \{0\}$), if the minimiser $\widetilde{p} \in \operatorname{int}\mathcal{K}$. Unfortunately this may not be so, unless the norm is small enough that there actually exists a solution $(p, a) \in \mathcal{C}_F(1)$. In some applications, as we shall see

in Section 4.5, the pseudo-inverse however provides a usable result (and is the solution for $\bullet = -\infty$ as well, in fact).

Remark 4.5. In the SCP method and case i) of Lemma 4.12, actually $O(\rho) = O(\Delta_0)$, so $\overline{\epsilon} = O(\Delta_0 + \|V_f\|_F \|c_f\|_F)$. This is because, if $f_\nu(y) - \min f_\nu \leq 2\rho \leq \underline{\epsilon}$, then choosing $z \in \partial\nu(y)$, we have $z \in \partial_{\underline{\epsilon}} f(y)$, by Lemma 4.9, so y is $\underline{\epsilon}$-semi-critical.

Remark 4.6. When the function ν can also be expressed as (3.7), we may actually use any feasible initialisation at $y_{[k]}$, by solving for $z \in \partial_\rho \nu(y_{[k]})$ simultaneously with $z \in \partial_{\epsilon'} f(\hat{y})$, keeping z free during the interior point method instead of choosing and initialising at one. This is particularly well suited with the modification in Remark 4.2 of Algorithm 4.3. For, replacing $B = (B_f; B_\nu)$ with $B' = (B_f; 0)$, in (4.2)–(4.4) and advancing in the direction of any solution to the system, will not alter the y component of d_ν. In fact, the components of the equation only featuring ν, become

$$A_\nu^* \Delta\lambda_\nu + \Delta d_\nu = 0,\ A_\nu \Delta p_\nu = 0,\ L(p_\nu)\Delta d_\nu + L(d_\nu)\Delta p_\nu = \Delta q_\nu.$$

This equation is fully determined in the interior of $\mathcal{K}_\nu$ under standard assumptions following from (3.8)–(3.10); see Faybusovich [1997b]. Therefore it would suffice to solve this equation first, and then use $\Delta z = B_\nu \Delta p_\nu$ for solving the f component separately. This argument also shows that the resulting full system of equations is non-singular.

However, the rate of decrease can be low. More precisely: We have $(p_\nu, d_\nu) \in \mathcal{C}_{-\infty}(\gamma')$ for some $\gamma' \leq (\gamma + r_f/r_\nu)/(1 + r_f/r_\nu) \in (0,1)$ when $(p,d) \in \mathcal{C}_{-\infty}(\gamma)$.[2] Therefore, since $\langle \Delta p_\nu, \Delta d_\nu \rangle = 0$, the results of Section 4.2.4 show that $(p_\nu(\alpha), d_\nu(\alpha)) \in \mathcal{C}_{-\infty}(\gamma')$ for $\alpha \in (0, \bar{\alpha})$ with $\bar{\alpha}$ bounded away from zero. Since $\alpha\Delta z$ is bounded by $\mathcal{R}(\partial\nu)$ being bounded, Δz must then also be bounded. However, since $\langle \Delta p_f, \Delta d_f \rangle = -\langle \Delta y, \Delta z \rangle \neq 0$ (generally), the condition of the system

$$B_f^* \Delta y + A_f^* \Delta\lambda_f + \Delta d_f = 0,\ A_f \Delta p_f = 0,\ B_f \Delta p_f = \Delta z,$$
$$L(p_f)\Delta d_f + L(d_f)\Delta p_f = \Delta q_f$$

for the f components, can limit the step length considerably. The convergence can therefore become slow, or even close to a halt as the iterates close to a point (with $\mu = 0$) that fails to be non-degenerate and strictly complementary (for f), as the system can be singular there.

4.4.3 Application of SCP to restoration phase

A variant of Algorithm 4.3 can be used for restoration in Algorithm 4.2, and it never fails, so that convergence is attained. We simply add after Step 4 (of Algorithm 4.3) the step:

[2] This can be seen by approximating $(1-\gamma)\sum_i \zeta_i(q)/(r_f + r_\nu) \leq \min_i\{\zeta_i(q)\} \iff \|P_e^\perp q\|_{-\infty} \leq \gamma\mu(q)$ from above and below to remove the eigenvalues of the q_f component of $q = (q_f, q_\nu) = Q_p^{1/2} d$.

4^+. Calculate (p, d, ϵ) such that $(p, d) \in G_\epsilon^{-1}(\hat{y}, 0) \cap \mathcal{C}_\bullet(\gamma)$ (for the data of f_ν). If $(f_\nu(\hat{y}), \epsilon)$ is acceptable to $\mathcal{F}$, return to the main phase with result $(p, d, \hat{y}, \epsilon)$.

If the basic version of Algorithm 4.3 is used (or the variant of Remark 4.2, but not that of Remark 4.6), then provided that $\bar{\epsilon}$ is large enough that the initialisation required by Theorem 4.3 can be performed (cf. Lemma 4.12), we have the bound $O(K_{\gamma_f, r_f} \tau_\rho^{-1} \log \tau_\rho^{-1})$ with $\tau_\rho \triangleq (\underline{\epsilon} - 2\rho)/\bar{\epsilon}$ for the number of interior point iterations in each restoration phase. Since $\tau_\rho \leq \tau = \underline{\epsilon}/\bar{\epsilon}$, the total number of interior point iterations in Algorithm 4.2 (with those in the main phase for f_ν, and those in the restoration phase for f alone), is therefore bounded by $O(K_{\gamma_f, r_f} \tau_\rho^{-2} (\log \tau_\rho^{-1})^3)$, provided that the conditions in Theorem 4.2 are satisfied, including $\epsilon \leq \bar{\epsilon}$ on return from Step 4^+ above.

This bounded reinitialisation in Step 4^+ can indeed be enforced by adding such a check (or including $(0, \bar{\epsilon})$ in the filter), in which case the SCP restoration method simply churns out new candidates while decreasing f_ν, until it reaches an $\underline{\epsilon}$-semi-critical point or an acceptable candidate. The check does not degrade the complexity bounds calculated above, because SCP alone has lower complexity. It is thus seen that the complexity of the method is entirely dependent on τ, the worst initialisation quality proportional to the desired solution quality, and ψ, which describes the proportion of the concave component and closeness to level-unboundedness of f_ν.

We may, however, also calculate some bounds for reinitialisation quality, to ensure that provided with big enough but reasonably bounded $\bar{\epsilon}$ and γ, the enforcement of $\epsilon \leq \bar{\epsilon}$ does not simply reduce the filter method to SCP. The next result proves the existence of such a "good" initialiser; later a more practical procedure is provided, with bounds not so directly related to the quality of the current iterate. Note from the proof that the bounds are also good for initialisation (of f data) for SCP restoration, in addition to reinitialisation (of f_ν data) on return to the primary phase.

Theorem 4.4. *Fix the constants $\underline{\epsilon} \geq 2\rho > 0$. Suppose Assumption 4.1 holds for f and $\mathcal{R}(\partial v) \subset \psi \mathcal{R}(\partial f)$ for some $\psi \in (0, 1)$. Suppose moreover that $f_\nu(y) - \min f_\nu \leq \Delta_0$. Then either of the following holds:*

(i) y is $\underline{\epsilon}$-semi-critical for f_ν.

(ii) There exists $(p, d) \in G_{\bar{\epsilon}}^{-1}(y, 0) \cap \mathcal{C}_{-\infty}(\gamma)$ for $\bar{\epsilon} = O(\Delta_0 + \|V_f\|_F \|c_f\|_F)$, $\gamma \in [0, 1)$ with $(1 - \gamma)^{-1} = O((1 - \psi)^{-2} \tau^{-1})$, and $\tau \triangleq \underline{\epsilon}/\bar{\epsilon}$.

Proof. Find $z \in \mathbb{R}^m$ and $(p_\nu, d_\nu) \in G_{\nu,\rho}^{-1}(y, z) \cap \mathcal{C}_{-\infty}(\psi)$ with *exactly* $\langle p_\nu, d_\nu \rangle = \rho$. This can be done, even with $\psi = 0$, because the selection $p_\nu \circ d_\nu = \mu_\nu e$, with $\mu_\nu \triangleq \rho / r_\nu$, within $\partial \nu$ comes from the subdifferential of a barrier-smoothed function; cf. Remark 3.1. An alternative way to see this, is to write $\xi_\nu \triangleq -B_\nu^* y - c_\nu$, to get the system of equations

$$A_\nu^* \lambda_\nu + d_\nu = \xi_\nu', \; A_\nu p_\nu = b_\nu, \; p_\nu \circ d_\nu = \mu_\nu e; \; p_\nu, d_\nu \in \mathcal{K}_\nu, \tag{4.19}$$

which [see, e.g., Faybusovich, 1997b; Schmieta and Alizadeh, 2003] characterises the solutions of

$$\min\left[\langle \xi_\nu, p_\nu\rangle - \mu_\nu \log(\det p_\nu)\right] \quad \text{subject to} \quad A_\nu p_\nu = b_\nu,\ p_\nu \in \mathcal{K}_\nu.$$

With z and (p_ν, d_ν) found, apply Lemma 4.12 to find $(p_f, d_f) \in G^{-1}_{f,\epsilon}(y,z) \cap \mathcal{C}_{-\infty}(\gamma')$ for some $\epsilon = \langle p_f, d_f\rangle = O(\Delta_0 + \rho + \|V_f\|_F\|c_f\|_F)$, and $\gamma' \in [0,1)$ with $(1-\gamma')^{-1} = O((1-\psi)^{-1})$. Apply the following Lemma 4.13, to get the claim of the theorem at y for $\epsilon = O(\bar{\epsilon}) \triangleq O(\Delta_0 + 2\rho + \|V_f\|_F\|c_f\|_F)$ and $\tau_y^{-1} = O(\tau^{-1})$. Finish the proof by referring to Remark 4.5 to take out ρ from the complexity. □

Lemma 4.13. *Assume we have fixed $\underline{\epsilon} \geq 2\rho \geq \theta\underline{\epsilon} > 0$ for some $\theta > 0$. Suppose that for some $\gamma' \in [0,1)$ and $\epsilon' > 0$, we have $(p_f, d_f) = G^{-1}_{f,\epsilon'}(y,z) \cap \mathcal{C}_{-\infty}(\gamma')$ and $(p_\nu, d_\nu) = G^{-1}_{f,\rho}(y,z) \cap \mathcal{C}_{-\infty}(\gamma')$ with* exactly *$\epsilon' = \langle p_f, d_f\rangle$ and $\rho = \langle p_\nu, d_\nu\rangle$. Then either of the following holds:*

(i) $\epsilon' + \rho \leq \underline{\epsilon}$, in which case y is $\underline{\epsilon}$-semi-critical for f_ν.

(ii) $(p,d) = ((p_f, p_\nu), (d_f, d_\nu)) \in G^{-1}_{\epsilon}(y,0) \cap \mathcal{C}_{-\infty}(\gamma)$ for $\epsilon \triangleq \epsilon' + \rho$, and $\gamma \in [0,1)$ with $(1-\gamma)^{-1} = O((1-\gamma')^{-2}\tau_y^{-1})$, and $\tau_y \triangleq \underline{\epsilon}/\epsilon$.

Proof. Let $q = (q_f, q_\nu) \triangleq (Q^{1/2}_{p_f} d_f, Q^{1/2}_{p_\nu} d_\nu) = Q^{1/2}_p d$. Denoting $\underline{\zeta}(q) \triangleq \min_i \zeta_i(q)$, we have

$$(1-\gamma')\mu(q_f) \leq \underline{\zeta}(q_f) \leq \mu(q_f), \tag{4.20}$$

and likewise for ν. Therefore

$$(1-\gamma')\mu(q) = (1-\gamma')\frac{r_f\mu(q_f) + r_\nu\mu(q_\nu)}{r_f + r_\nu} \leq \frac{r_f\underline{\zeta}(q_f) + r_\nu\underline{\zeta}(q_\nu)}{r_f + r_\nu}.$$

But

$$r_\nu\underline{\zeta}(q_\nu)/\underline{\zeta}(q_f) \leq r_\nu\frac{\mu(q_\nu)}{\mu(q_f)}/(1-\gamma') \leq r_f\frac{\rho}{\epsilon'}/(1-\gamma'),$$

employing (4.20), as well as the exactness assumption in the denominator estimate. Because an analogous estimate holds with the roles of f and ν reversed, and $\underline{\zeta}(q) = \min\{\underline{\zeta}(q_f), \underline{\zeta}(q_\nu)\}$, we have

$$(1-\gamma')\mu(q) \leq \frac{1 + \max\{\epsilon'/\rho, \rho/\epsilon'\}}{1-\gamma'}\underline{\zeta}(q).$$

If $\epsilon' \leq \rho$, then $\epsilon' + \rho \leq 2\rho \leq \underline{\epsilon}$, which is covered by case (i). So assume the contrary.

We now get $1 + \epsilon'/\rho = (\rho + \epsilon')/\rho \leq \epsilon/(\theta\underline{\epsilon})$. Therefore, with γ defined by $(1-\gamma)^{-1} = (1+\epsilon'/\rho)(1-\gamma')^{-2}$, we have $(1-\gamma)^{-1} = O((1-\gamma')^{-2}\tau_y^{-1})$, as well as $\gamma \in [0,1)$ and $(1-\gamma)\mu(q) \leq \underline{\zeta}(q)$. Hence, case (ii) applies. □

We can in principle solve (4.19) approximately by standard interior point methods. After all, instead of $p_\nu \circ d_\nu = \mu_\nu e \in \mathcal{C}_{-\infty}(0)$, we only wanted $\mathcal{C}_{-\infty}(\psi)$. Then we could calculate (p_f, d_f) and modify the result as indicated in the proof. However, we would have to bound the quality of the initialisation for this method, which would annoyingly seem to involve y or the linearisation error e_ν (defined in Section 2.2). Sometimes (4.19) can be solved directly, however, as the examples below show. After that, we would still have to find $(p_f, d_f) \in G^{-1}_{f,\epsilon'}(y,z) \cap \mathcal{C}_{-\infty}(\gamma')$ as discussed towards the end of Section 4.4.2 above.

Example 4.2. Suppose Assumption 4.1 holds for ν. Let $\xi_\nu = -B^*_\nu y - c_\nu$ be as in Theorem 4.4. Then, dropping the ν-subscripts to simplify the notation for this example, $d_i = \xi_i + \lambda_i a'_i$ and $p_i = \mu d_i^{-1}$, assuming λ_i is big enough for d_i to be invertible. The problem now is to have $\langle a'_i, p_i \rangle = b_i$, i.e., $\operatorname{tr}(Q^{1/2}_{a'_i} d_i^{-1}) = b_i/\mu$. Taking $Q^{1/2}_{a'_i}$ inside d_i^{-1} [doable by, e.g., Faraut and Korányi, 1994, Proposition II.3.3],

$$\operatorname{tr}(Q^{-1/2}_{a'_i}\xi_i + \lambda_i e)^{-1} = b_i/\mu. \tag{4.21}$$

Thus, if we can invert the trace of the *resolvents* of $\underset{\sim}{\xi}_i \triangleq Q^{-1/2}_{a'_i}\xi_i$, we can solve (4.19).

Example 4.3. Suppose that (each) $\mathcal{K}_i$ in Example 4.2 is a second order cone. Then for $x = (x^0, \bar{x})$, we have $x^{-1} = (x^0, -\bar{x})/\det(x)$, $\det(x) = (x^0)^2 - \|\bar{x}\|^2$, and $\operatorname{tr} x = \langle e, x \rangle = 2x^0$. By Assumption 4.1, $\langle e, \underset{\sim}{\xi}_i \rangle = \langle (a'_i)^{-1}, \xi_i \rangle = 0$, which implies $\underset{\sim}{\xi}^0_i = 0$. Therefore, $\operatorname{tr}(\underset{\sim}{\xi}_i + \lambda_i e)^{-1} = 2\lambda_i/(\lambda_i^2 - \|\underset{\sim}{\bar{\xi}}_i\|^2)$, so we get from (4.21) the quadratic equation $(2\mu_\nu/b_i)\lambda_i = \lambda_i^2 - \|\underset{\sim}{\bar{\xi}}_i\|^2$. This can be solved for λ_i, as we wanted.

The proof of the next result provides a simpler practical reinitialisation method, which has worse bounds near an actual minimum of f_ν. It however appears to work better in practise, and provide lower ϵ, which may be attributable to the fact the method is seldom near the global minimum, but rather a local one or some other semi-critical point. The subgradient assumptions are guaranteed by the SCP procedure.

Lemma 4.14. *Suppose Assumption 4.1 holds (for both f and ν), and that $z \in \partial f_{\epsilon'}(\hat{y})$ and $z \in \partial_\rho \nu(y)$. Denote the linearisation error of ν by $\ell \triangleq e_\nu(\hat{y}; y, z)$. Then for all $\psi \in (0,1)$, there exist $(p,d) \in G^{-1}_\epsilon(\hat{y}, 0) \cap \mathcal{C}_{-\infty}(\gamma)$ with $\gamma \triangleq (1+\psi)/2$ and $\epsilon/2 \triangleq \psi(\epsilon' + \rho + \ell) + (1-\psi)(f(\hat{y}) + \nu(\hat{y}))$.*

Proof. We note that by the definition of f, there exists $\hat{p}_f(y) \in V_f$ such that $f(y) = \langle B^*_f y + c_f, \hat{p}_f(y) \rangle$. Furthermore, by (3.11), there exists $\hat{d}_f(y) = -B^*_f y - c_f - A^*_f \hat{\lambda}_f(y) \in \mathcal{K}_f$ such that $\langle \hat{p}_f(y), \hat{d}_f(y) \rangle = 0$. Therefore, for all $p'_f \in V_f$,

$$f(y) - \langle B^*_f y + c_f, p'_f \rangle = \langle \hat{d}_f(y), \hat{p}_f(y) - p'_f \rangle = \langle \hat{d}_f(y), p'_f \rangle. \tag{4.22}$$

An analogous result holds for ν.

By the approximate subgradient transportation formula (see Section 2.2), $z \in \partial\nu_{\rho+\ell}(\hat{y})$. Therefore, we can find $(p', d') \in G^{-1}_{\epsilon'+\rho+\ell}(\hat{y}, 0)$. In fact, we can take $d' = \hat{d} \triangleq (\hat{d}_f(\hat{y}), \hat{d}_\nu(\hat{y}))$, since with $\hat{y}$ fixed, the choice $\hat{\lambda}(\hat{y})$ for λ must minimise $d' \mapsto \langle d', p' \rangle$. (If some other d' at y achieved lower value, then also $\langle d', \hat{p} \rangle < \langle \hat{d}, \hat{p} \rangle = 0$, which is a contradiction to properties of symmetric cones.) We therefore have by Assumption 4.1 and (4.22) with $p' = a^{-1}$ that

$$\langle a^{-1}, d' \rangle = f(\hat{y}) + \nu(\hat{y}) - \langle B^*\hat{y} + c, a^{-1} \rangle = f(\hat{y}) + \nu(\hat{y}). \tag{4.23}$$

Now we simply apply Lemma 4.11 with $\gamma' = 1$ and $\gamma = (1 + \psi)/2$ to yield the claim for

$$\epsilon/2 = \psi\langle p', d' \rangle + (1 - \psi)\langle a^{-1}, d' \rangle.$$

Then we just use (4.23) and $\langle p', d' \rangle \leq \epsilon' + \rho + \ell$. □

Remark 4.7. The subgradient transportation formula actually holds for fixed p'_ν. To see this, suppose $(p'_\nu, d'_\nu) \in G^{-1}_{\nu,\rho}(y, z)$ and calculate $\langle p'_\nu, \hat{d}_\nu(\hat{y}) \rangle = \langle p'_\nu, \hat{d}_\nu(y) \rangle + \langle p'_\nu, \hat{d}_\nu(\hat{y}) - \hat{d}_\nu(y) \rangle = \langle p'_\nu, \hat{d}_\nu(y) \rangle + \nu(\hat{y}) - \nu(y) - \langle B^*_\nu(\hat{y} - y), p'_\nu \rangle \leq \rho + \ell$, where we have applied (4.22) twice in the last equality.

What this means is that we can with simple modifications of (p'_ν, d'_ν) and $(p'_f, d'_f) \in G^{-1}_{f,\epsilon'}(\hat{y}, z)$, produce (p, d) satisfying the claims of Lemma 4.14: calculate $\hat{d}(\hat{y})$, translate $p' = (p'_f, p'_\nu)$ towards a^{-1} by $1 - \psi$, and add a factor of a to $\hat{d}(\hat{y})$.

Remark 4.8. As we see, to ensure that $(p, d) \in \mathcal{C}_{-\infty}(\gamma)$, without any further knowledge of the containment in $\mathcal{C}_\bullet(\gamma')$ of $(p'_\nu, d'_\nu) \in G^{-1}_{\nu,\rho}(\hat{y}, z)$ after transportation of z from y to $\hat{y}$, we have to ensure that p is also far enough from the boundary of $\mathcal{K}$. To do so, we apply the translation towards a^{-1}. But this component brings the annoying $f + \nu$ sum (instead of difference) into the bound, which is not found in the bound of Theorem 4.4.

4.5 Practical considerations and experience

We have not tested (and compared against other methods) our algorithms to any statistical significance. In this section, we however list some observations from our limited experience with the methods. But we begin with a note of another kind.

4.5.1 Reductions of the linear system

The system (4.2)–(4.4) can be huge. In a typical application to sums of K Euclidean distances in $\mathbb{R}^m$, the (block-diagonal) matrix A has size $(m + 1)K \times K$. Fortunately, we can reduce the system to only depend on the dimension of y.

Denote $F \triangleq L(d)^{-1}L(p)$. Then, multiplying (4.4) from the left by $L(d)^{-1}$ and expanding Δd from (4.3), the system becomes

$$(A, B_-)\Delta p = 0,\ -F(A,B)^*(\Delta\lambda, \Delta y) + \Delta p = L(d)^{-1}\Delta q.$$

Multiplying the second equation by (A, B_-), we get the *normal equations*, standard in interior point methods,

$$(A, B_-)F(A,B)^*(\Delta\lambda, \Delta y) = -(A, B_-)L(d)^{-1}\Delta q, \tag{4.24}$$

The first block of lines says

$$AFA^*\Delta\lambda = -Au \triangleq -A(FB^*\Delta y + L(d)^{-1}\Delta q).$$

Now, AFA^* is positive-definite when F is symmetric positive-definite (ensured by operator commutative scaling when $p, d \in \operatorname{int}\mathcal{K}$), for we have assumed $\mathcal{R}(A^*)$ to be full. In fact, when the product-form representation of Section 3.4.5 holds, AFA^* is a positive definite diagonal matrix. Denote $X \triangleq A^*(AFA^*)^{-1}A$. Then $A^*\Delta\lambda = -Xu$. The second block of lines from (4.24) says now

$$\begin{aligned}
& B_-F(A,B)^*(\Delta\lambda, \Delta y) = -B_-L(d)^{-1}\Delta q \\
\iff & -B_-FXu + B_-FB^*\Delta y = -B_-L(d)^{-1}\Delta q \\
\iff & B_-(F - FXF)B^*\Delta y = -B_-(I - FX)L(d)^{-1}\Delta q \\
\iff & B_-\tilde{F}B^*\Delta y = -B_-\tilde{F}L(p)^{-1}\Delta q,
\end{aligned}$$

with $\tilde{F} \triangleq F - FXF$. It thus suffices to solve this reduced equation, again of standard normal equation form. Note that when $\Delta q = \sigma\mu e - p \circ d$, $L(p)^{-1}\Delta q = \sigma\mu p^{-1} - d$. We may therefore (still) incorporate all the effects of scaling into $\tilde{F}$: $\underset{\sim}{B}_-\tilde{F}\underset{\sim}{B}^* = B_-(Q_v^{-1}\tilde{F}Q_v^{-1})B^*$ and $\underset{\sim}{B}_-\tilde{F}(\sigma\mu\tilde{p}^{-1} - \underset{\sim}{d}) = B_-(Q_v^{-1}\tilde{F}Q_v^{-1})(\sigma\mu p^{-1} - d)$, when $\tilde{F}$ already uses scaled data.

Example 4.4 (Sums of Euclidean norms)**.** Recall from Example 3.3 that in this case $A(p_1, \ldots, p_n) = (\langle e, p_1\rangle, \ldots, \langle e, p_n\rangle)$ with $p_i = (p_i^0, \bar{p}_i)$ in a second order cone. That is, $A = \operatorname{diag}(\langle e, \cdot\rangle, \ldots, \langle e, \cdot\rangle)$, and $A^* \propto \operatorname{diag}(e, \ldots, e)$. Therefore $AFA^* \propto \operatorname{diag}(\langle d_1^{-1}, p_1\rangle, \ldots, \langle d_n^{-1}, p_n\rangle)$ and X is block-diagonal with $1/\langle d_i^{-1}, p_i\rangle$ in the top-left corner of each block and zero elsewhere.

4.5.2 Various practical remarks and examples

Remark 4.9 (Initialisation for SCP restore)**.** If we use the variant of the SCP method discussed in Remark 4.6, simultaneously solving for $z \in \partial_{\epsilon'} f(\hat{y})$ and $z \in \partial_\rho \nu(y_{[0]})$, we will be able to directly use (p, d, y) from the main filter interior point method (Algorithm 4.2) in the first iteration of the restoration method, assuming the same neighbourhood $\mathcal{C}_\bullet(\gamma)$ is used. Therefore no specific initialisation is needed there. However, between iterations of the SCP method (if multiple iterations are needed), and on return from it, reinitialisation is needed to update d_ν to reflect $\hat{y}$, and to maintain constraints on the γ-neighbourhood.

Remark 4.10 (Neighbourhoods)**.** Since Algorithm 4.2 sets an explicit bound on the minimum decrease in ϵ, and uses the SCP method otherwise, we do not particularly need the bounds in the decrease that Lemmas 4.5 and 4.6 provide for the neighbourhoods $\mathcal{C}_\bullet$ with $\bullet \in \{F, 2, -\infty\}$ – and which blow up near singularities in our non-convex case. Thus it seems beneficial to use other neighbourhoods that are easier to initialise to provide small ϵ. If $\mathcal{K} = \prod \mathcal{K}_i$ for smaller symmetric cones $\mathcal{K}_i$, one could therefore consider the neighbourhood defined as (topological) product of the neighbourhoods $\mathcal{C}_\bullet$ for each $\mathcal{K}_i$. In this case we also use as Δq, not $\sigma\mu e - p \circ d$, but the product of these for each sub-algebra (i.e., different μ for each component).

In the restoration phase we want to use the standard neighbourhoods and Δq to ensure convergence. The initialisation optimisation of Remark 4.9 is, however, partially lost when product neighbourhoods are used in the main phase. The problem is that although the eigenvalues of q_i for $q = p \circ d$ are bounded away from zero proportionally to $\mu(q_i)$, this value may itself become small in relation to the overall $\mu(q)$, and therefore even for $\mathcal{C}_{-\infty}(\gamma')$, the required γ' may approach 1. It is, of course, possible to set a bound on γ', and do complete reinitialisation, if it is violated.

Example 4.5 (SCP initialisation for spatial medians)**.** Consider again the problem from Lemma 4.12, of solving $B_f p_f = z/\psi$ with $p_f \in V_f$. In the simple case of the spatial median in $\mathbb{R}^m$, $f(y) = \sum_{i=1}^n \|y - a_i\|$, as in general for sums of Euclidean norms, we have $p_f = (p_{f,1}, \dots, p_{f,n})$ with $p_{f,i} = (p^0_{f,i}, \bar{p}_{f,i}) \in \mathcal{E}^{m+1}$, and $a_i' = e$. Furthermore, $B_f p_f = \sum_i \bar{p}_{f,i}$, so that a simple solution with $p_{f,i} = p_{f,j}$ exists, when at all $z \in B_f V_f$. This obviously extends to sums of spatial medians ($\sum_k f(y_k)$), and suffices for our forthcoming application examples, where $\mathcal{R}(\partial\nu)$ is small enough to be covered by the spatial median component of f, and we may therefore take $\bar{p}_{f,i} = 0$ for any remaining terms.

4.5.3 Application to a clustering formulation

The primary applications we had in mind in the study Algorithms 4.1 and 4.2, was the MO clustering formulation studied in Chapter 6, as well as the MO-TSP formulation of the Euclidean TSP studied in Chapter 7. The former, as already seen in Chapter 1, reads with the notation $\bar{y} = (y_1, \dots, y_K) \in \mathbb{R}^{Km}$, $\bar{a} = (a_1, \dots, a_n) \in \mathbb{R}^{nm}$ as

$$\min_{\bar{y}} f(\bar{y}; \bar{a}) - w\nu_{\mathrm{MO}}(\bar{y}) \tag{4.25}$$

for some $w \in (0, n/(K-1))$, and

$$f(\bar{y}; \bar{a}) \triangleq \sum_{i=1}^{K} \sum_{k=1}^{n} \|y_i - a_k\|, \quad \nu_{\mathrm{MO}}(\bar{y}) \triangleq \sum_{i<j} \|y_i - y_j\|.$$

In the latter problem, we set $K = n$, $w = 1$, and add to (4.25), the path-length penalty $\lambda_{\mathrm{TSP}} f_{\mathrm{TSP}}(\bar{y})$ for some $\lambda_{\mathrm{TSP}} > 0$ and $f_{\mathrm{TSP}}(\bar{y}) \triangleq \sum_{i=1}^{n} \|y_i - y_{i+1}\|$ (with the identification $y_{n+1} = y_1$).

According to Chapter 6, $n(K-1)^{-1}\mathcal{R}(\partial v) \subset \mathcal{R}(\partial f)$. Therefore, by our choice of $w = 1$ in the TSP problem, we may take $\psi = (n-1)/n$ in Lemma 4.12 and Theorem 4.4, and obtain $(1-\gamma)^{-1} = O((1-\psi)^{-1}) = O(n)$. Thus the complexity of the method in this application only depends polynomially on n (through both $r = 2(n^2 + n(n-1)/2 + n) = 3n^2 + n$ and γ), and log-polynomially on the reciprocal of the desired relative solution quality τ. Recall from Example 4.5 that as a sum of spatial medians, finding $p_f \in V_f$ satisfying $B_f p_f = z$ is easy, while we may choose $B_{\text{TSP}} p_{\text{TSP}} = 0$ ($p_{\text{TSP}} = e$) for the f_{TSP} component.

Our principal practical observations from application of Algorithm 4.2 to these problems are as follows:

(i) The spatial median of the data $\bar{a}$ is highly attractive: Unless the filter is initialised to forbid convergence to this point (by suitable values of $\overline{g}$), or if we can initialise the method with p and d (and not just y) close to some other attractor (semi-critical point, cf. Lemma 3.6), it is likely that many of the variables y_i will converge to the spatial median. Especially this appears to be a problem when K is a considerable proportion of n, such as in the MO-TSP case. However, strict initialisation of the filter may provoke long runs of the comparably slow SCP restoration method.

(ii) The reinitialisation method of Example 4.3, although with better theoretical bounds (Theorem 4.4), does not work so well in practise, as the method of Remark 4.7 (and Lemma 4.14).

(iii) In general, the performance is unpredictable: sometimes convergence is fast, and sometimes it becomes slow, spending a lot of time in the restoration phase. (Pure SCP by contrast provides slow but more consistent performance.)

(iv) So far it appears that an extension of the Weiszfeld method, which we study in the following chapter, provides more consistent practical performance. At least in the MO-TSP application, our experience is that it provides reasonable results in fewer iterations.

In summary, we find that although the theoretical basis of our method is sound, more research and experimentation is still needed to find out if and with what parametrisation and modifications, the algorithm can provide competitive practical performance in these, and other, applications. Such practically-oriented study is outside the scope of this mainly theoretical thesis.

5 THE WEISZFELD METHOD AND PERTURBED SPATIAL MEDIANS

5.1 Introduction

In this chapter, we are interested in minimisation problems, where the objective function can be modelled as a perturbed version of the objective function for the spatial median. More specifically, what concerns us are diff-convex problems of the form

$$\min_y \left(\sum_{k=1}^{n} \|W_k(a_k - y)\| - \nu(y) \right) \tag{5.1}$$

for some fixed points $a_1, \ldots, a_n$, weight matrices W_k, and a convex function ν.

On the application side, which we will leave to Chapter 6, we are primarily concerned with location problems involving multiple prototypes to be placed according to some optimality criterion that can be given the form (5.1). This problem seems, at a first glance, to only involve a single prototype. However, suitably choosing the matrices W_k to model incomplete data, will let us model multi-prototype problems as single-prototype ones.

Indeed, the algorithm we develop for problems of the form (5.1), will be a further extension of the generalisation to incomplete data sets of the Weiszfeld algorithm in Kärkkäinen and Äyrämö [2004, 2005] and Valkonen [2006, 2008a]. This algorithm in its basic form [Weiszfeld, 1937; Kuhn, 1973] seeks a minimiser to $\sum_{i=1}^{n} w_i d(a_i, \cdot)$ for the Euclidean distance in $\mathbb{R}^m$ by iterating

$$T : y \mapsto \frac{\sum_{i=1}^{n} s_i a_i}{\sum_{i=1}^{n} s_i} \quad \text{with } s_i = w_i / d(a_i, y). \tag{5.2}$$

Since the objective of (5.1) is a difference of convex functions, 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, as seen in the following chapters.

This chapter is organised as follows. First, in Section 5.2, we introduce the problem of the perturbed spatial median in more detail. Then, in Section 5.3, we study directions of descent for this objective function. Based on these results, along with studying optimality conditions, we define the perturbed Weiszfeld method in Section 5.4. Its convergence is then studied in Section 5.5. We conclude the chapter with an analysis of $\mathcal{R}(\partial f)$ in Section 5.6, useful for obtaining level-boundedness in the applications of the following Chapters 6 and 7.

5.2 The perturbed spatial median

Throughout most of this chapter, we work with $n \geq 1$ vertices $a_1, \ldots, a_n \in \mathbb{R}^m$ and diagonal positive-semidefinite matrices $W_1, \ldots, W_n \in \mathbb{R}^{m \times m}$. 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 seminorms and distance functions $d_i(y) \triangleq \|a_i - y\|_i \triangleq \|W_i(a_i - y)\|$ as well, and the sum of distances function $f : \mathbb{R}^m \to \mathbb{R}$ as

$$f(y) \triangleq \sum_{i=1}^n d_i(y) = \sum_{i=1}^n \|W_i(a_i - y)\|. \tag{5.3}$$

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 [Valkonen, 2006, Theorem 3.1], where the problem (5.3) was studied under a more elaborate model for missing data.

Now, consider the problem of finding the minimum of (5.3) 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

$$\min_{y \in \mathbb{R}^m} f_\nu(y) = \min_{y \in \mathbb{R}^m} \sum_{i=1}^n d_i(y) - \nu(y). \tag{5.4}$$

Any solution of problem (5.4) 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., [Rockafellar, 1972, 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.

5.3 Directions of descent

Notice that, since ν is convex, if we replace it with a linearisation $\tilde{v}_y^v(y') \triangleq \nu(y) + v^T(y' - y)$ for $v \in \partial\nu(y)$, then $-\nu \leq -\tilde{v}_y^v$ and, furthermore, f_ν is dominated by the *upper convexification* $f_{\tilde{v}_y^v}$. Therefore, for any $y' \in \mathbb{R}^m$, for which $f(y') - \tilde{v}_y^v(y') < f(y) - \tilde{v}_y^v(y) = f(y) - \nu(y)$ it follows that $f(y') - \nu(y') < f(y) - \nu(y)$. This means that if some upper convexification at y 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 $\pi(y) \triangleq \{i \mid W_i(a_i - y) = 0\}$. The gradient of the differentiable components of f at y is then given by

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

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

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

Theorem 5.1. *Suppose $\nu(y) = v^T y$ for some $v \in \mathbb{R}^m$, and let $z \in \mathbb{R}^m$. Then $f_\nu(y + \omega z) < f_\nu(y)$, if $\omega \in (0, \Omega)$ with $\Omega \triangleq \Omega(y, v, z)$ defined as the supremum of ω' satisfying*

$$\omega'(z^T S_\pi(y) z) < -2h(z, v; y). \tag{5.5}$$

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

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

Write

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

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

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

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

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

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

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

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 (5.5).

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

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

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

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

Lemma 5.1. *Let v and f_ν be as in Theorem 5.1. Let $\tilde{z} \in Z(y)$ be such that $h(\tilde{z}, v; y) < 0$ if such a choice exists. Otherwise choose $\tilde{z} = 0$. Suppose $\omega \in (0, 2)$ and that $y \in \mathbb{R}^m$ is not a minimiser of f_ν. Then*

$$z = z(y, v) \triangleq -\bar{\rho}_{\pi(y)} S_\pi^\dagger(y) g_\pi^v(y) + \alpha \tilde{z} \tag{5.6}$$

is a direction of descent for f_ν when $\alpha \in (0, \alpha_0)$, where $\alpha_0 \triangleq \alpha_0(\omega, \tilde{z}, v; y) > 0$ is the supremum of α for which z satisfies the condition (5.5) *at y for given ω and v. Furthermore, $\alpha_0(2, \tilde{z}, v; y)$ gives for any $\omega \in (0, 2)$ a lower bound $\alpha_2(\tilde{z}, v; y) \in (0, \alpha_0]$ (strict if $-\bar{\rho}_{\pi(y)} g_\pi^v(y) \neq 0$), obtained as the supremum of α satisfying*

$$\alpha \tilde{z}^T \rho_{\pi(y)} S_\pi(y) \tilde{z} \leq -h(\tilde{z}, v; y). \tag{5.7}$$

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

Inserting (5.6) into condition (5.5) of Theorem 5.1, we get

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

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

$$\omega((g_\pi^v)^T \bar{\rho}_\pi S_\pi^\dagger g_\pi^v + \alpha^2 \tilde{z}^T \rho_\pi S_\pi \tilde{z}) - 2((g_\pi^v)^T \bar{\rho}_\pi S_\pi^\dagger g_\pi^v - \alpha h(\tilde{z}, v; y)) < 0, \qquad (5.8)$$

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

If $\tilde{z} = 0$, then α does not contribute to (5.8), so its choice is irrelevant and α_0 infinite. If, furthermore, $\bar{\rho}_\pi g_\pi^v = 0$, then $\min h(z, v; y) = 0$ over $\|z\| = 1$, and therefore by Theorem 5.1, y is a minimiser of f_v. 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 $h(\tilde{z}, v; y) < 0$, 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 (5.8) 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 $\alpha_0(\omega, \tilde{z}, v; y) > 0$, such that $\alpha \in (0, \alpha_0(\omega, p, v))$ satisfies the inequality.

Setting $\omega = 2$ in (5.8), gives the condition for α_2. Furthermore, if α_2 satisfies (5.8) 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 5.1.

(i) When $\pi(y) = \emptyset$, necessarily $\tilde{z} = 0$, and we get from (5.6) that $z(y, v) = -S_\pi^\dagger(y) g_\pi^v(y)$. If $W_k = w_k I$, i.e., the weights are uniform and no data is missing, $S_i = w_i I / \|y - a_i\|$, and this step reduces to the the conventional Weiszfeld step used in (5.2).

(ii) When $\pi(y) = \{k\}$ is a singleton, a $\tilde{z}$ may be easily found by minimising $h(z, v; y) = g_\pi^v(y)^T z + \|z\|_k$ 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 $Z(y)$, so that we may choose $\tilde{z} = 0$ in this case. The result is therefore $\tilde{z} = -(W_k^\dagger)^2 g_\pi^v(y) / \|W_k^\dagger g_\pi^v(y)\| \in Z(y)$ when $\|W_k^\dagger g_\pi^v(y)\| \geq 1$, and $\tilde{z} = 0$ otherwise.

(iii) When $\#\pi(y) > 1$, but the data do not overlap, i.e., $\mathcal{R}(W_i) \cap \mathcal{R}(W_j) = \{0\}$ for distinct $i, j \in \pi(y)$, $\tilde{z}$ 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 $\#\pi(y) > 1$, 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 2 in any case establishes relevant formulae for the non-partially-overlapping/hierarchical case.

5.4 Optimality conditions and the method

We denote the set of semi-critical points for our problem of interest by P_∂. Recall that this means that $0 \in \partial f(y) - \partial\nu(y)$. But then, for some $v \in \partial\nu(y)$, the convex function $f - v^T : y \mapsto f(y) - v^T y$ (and then $f_{\tilde{v}_y^v}$) has minimum at y. We therefore find by Theorem 5.1 that y is semi-critical if and only if $h(z, v; y) \geq 0$ for all $z \in \mathbb{R}^m \setminus \{0\}$ for some (fixed) $v \in \partial\nu(y)$.

Recalling that semi-criticality is equivalent to criticality in the sense of Clarke subdifferentials, $0 \in \partial^\circ f_\nu(y)$ in particular when f is differentiable, we find that this is the case whenever $\pi(y) = \emptyset$. On the other hand, if some upper convexification of f_ν by $\tilde{v}_y^v$ does not have a minimum at y, 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 [Rockafellar and Wets, 1998], if $y_i \to y$ and $v_i \in F(y_i)$, imply that every accumulation point of $\{v_i\}$ is in $F(y)$.

Definition 5.1. Let $\mathcal{D}\nu$ be an outer-semicontinuous mapping, such that $\emptyset \neq \mathcal{D}\nu(y) \subset \partial\nu(y)$, for $y \in \mathbb{R}^m$. If $\partial f(y) \cap \mathcal{D}\nu(y) \neq \emptyset$, we refer to y 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 5.1, $\mathcal{D}$-criticality is equivalent to $h(z, v; y) \geq 0$ holding for all z for some $v \in \mathcal{D}\nu(y)$. 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

$$\mathcal{D}_N\nu(y) \triangleq \{\lim_{r\to\infty} \nabla\nu(y_{[r]}) \mid y_{[r]} \to p, \nu \text{ is differentiable at } y_{[r]}\},$$

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

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

Algorithm 5.1 (The perturbed SOR-Weiszfeld method)**.**

1. Set $r = 0$, and choose an initial iterate $y_{[0]} \in \mathbb{R}^m$. Choose $\mathcal{D}\nu$ satisfying Definition 5.1 (typically $\partial\nu$ or $\mathcal{D}_N\nu$), as well as a stopping criterion.
2. Choose $v_{[r]} \in \mathcal{D}\nu(y_{[r]})$, $\tilde{z} \in Z(y_{[r]})$, $\omega \in (1, 2)$ and $\alpha \in (0, \alpha_0(\omega, \tilde{z}, v_{[r]}; y_{[r]}))$, as described in Lemma 5.1.
3. Calculate $y_{[r+1]} \triangleq T_\omega(y_{[r]}, v_{[r]})$ with z defined by (5.6), and
$$T_\omega(y, v) \triangleq y + \omega z(y, v).$$
4. If the stopping criterion is not satisfied, continue from Step 2 with $r \triangleq r + 1$.

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

Lemma 5.2. *The iteration T_ω is descending for f_ν if $y \notin P_\mathcal{D}$.*

Proof. By Lemma 5.1, $z(y, v)$ is a direction of descend for $f_{\tilde{v}_y^v}$ when $v \notin \partial f(y)$, and therefore for f_ν as well. □

5.5 Convergence

We now turn to the convergence properties. The following Lemma 5.3 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 5.2, assuming the iterates do not diverge.

We denote $y' \triangleq T_\omega(y, v)$. We will sometimes omit v from the parameters for brevity, and write $\tilde{z}(y)$ etc. The specific selection is denoted $v(y)$.

Lemma 5.3. *Let the points and subgradients $y_{[r]} \in \mathbb{R}^m, v_{[r]} \in \mathcal{D}\nu(y_{[r]})$ $(r = 1, 2, \ldots)$ and $q \in \mathbb{R}^m, u \in \mathcal{D}\nu(q)$ be given, with constant $\pi' \triangleq \pi(y_{[r]}) \subsetneq \pi(q)$. Suppose that $\tilde{z} \in Z(q)$ with (i) $\rho_{\pi'}\tilde{z} = 0$, and (ii) $h(\tilde{z}, u; q) < 0$. If $(y_{[r]}, v_{[r]})$ 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*

$$\limsup_{r\to\infty} \frac{d_k(y'_{[r]})}{d_k(y_{[r]})} > 1. \tag{5.9}$$

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

Proof. Denote $y = y_{[r]}$ and $v = v_{[r]}$ for arbitrary r, for lighter notation. We may write

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

Since $\bar{\rho}_{\pi'} S_\pi^\dagger(y) S_\pi(y) = \bar{\rho}_{\pi'}$ by our prevailing assumption $\sum \mathcal{R}(W_k) = \mathbb{R}^m$, as well as $\rho_{\pi'}(y - q) = 0$, we have according to (5.6) that

$$\bar{\rho}_{\pi'}(y' - q) = y - q - \omega\bar{\rho}_{\pi'} S_\pi^\dagger(y) g_\pi^v(y) = (1 - \omega)(y - q) - \omega\bar{\rho}_{\pi'} S_\pi^\dagger(y)\tilde{g}^{v(y)}(y), \tag{5.10}$$

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

Let now $k \in \pi_{\tilde{z}}$. Since $W_k q = W_k a_k$, (5.9) follows if

$$\limsup_{r\to\infty} \frac{\|\bar{\rho}_{\pi'}(y'_{[r]} - q)\|_k}{\|y_{[r]} - q)\|_k} > 1.$$

Thus, by applying $\omega \geq 1$ and the reverse triangle inequality to the W_k-norm of (5.10), it becomes sufficient to show that for some k, $\limsup_{r\to\infty}(\omega N_k(y_{[r]}) - |1-\omega|) > 1$, i.e., $\limsup_{r\to\infty} N_k(y_{[r]}) > 1$, where

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

Suppose $\limsup_r N_k(y_{[r]}) \leq 1$ 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

$$\begin{aligned} \epsilon + \sum_{k\in\pi(q)\setminus\pi'} \|\tilde{z}\|_k &\geq \sum_{k\in\pi(q)\setminus\pi'} \|\tilde{z}\|_k N_k(y_{[r]}) \\ &\geq - \sum_{k\in\pi(q)\setminus\pi'} \tilde{z}^T W_k^2 \left(\frac{\bar{\rho}_{\pi'}}{d_k(y_{[r]})}\right) \left(\sum_{i\notin\pi'} \frac{\bar{\rho}_{\pi'} W_i^2}{d_i(y_{[r]})} \right)^{\dagger} \tilde{g}^{v[r]}(y_{[r]}) \qquad (5.11) \\ &= -\tilde{z}^T \left(\sum_{k\in\pi(q)\setminus\pi'} \Gamma_k(y_{[r]}) \right) \left(\sum_{i\notin\pi'} \Gamma_i(y_{[r]}) \right)^{\dagger} \tilde{g}^{v[r]}(y_{[r]}), \end{aligned}$$

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

Observe that $\Gamma_i(y_{[r]}) \to 0$ for $i \notin \pi(q)$, faster than for $i \in \pi(q) \setminus \pi'$ (if such were to happen). Therefore $\sum_{i\notin\pi'} \Gamma_i(y_{[r]}) - \sum_{k\in\pi(q)\setminus\pi'} \Gamma_k(y_{[r]}) \to 0$, and likewise for the pseudo-inverses. Now letting $\epsilon \searrow 0$ and going to the limit in (5.11) 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 5.4. *Suppose $(y_{[r]}, v_{[r]}) \to (q, u)$ with constant $\pi(y_{[r]}) = \pi'$ and $\tilde{z}(y_{[r]}) = 0$. Then we may take $\rho_{\pi'}\tilde{z}(q) = 0$.*

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

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

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

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

with the inequality holding by $\tilde{z}(y_{[r]}) = 0$. 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 5.1. The set of iterates $\{y_{[r]} \mid r = 1, 2, \ldots\}$ generated by Algorithm 5.1 is bounded. The function f_ν is bounded from below. The step sizes ω_r satisfy the conditions of Algorithm 5.1, and there exists $\overline{\omega} < 2$, such that $\omega_r \in [1, \overline{\omega}]$. Furthermore, $\tilde{z}(y_{[r]}) = 0$ (i.e., $\pi(y'_{[r]}) \supset \pi(y_{[r]})$) eventually.

Conditions ensuring level-boundedness of f, and hence the assumption on the boundedness of the iterates, will be studied in the following Section 5.6.

Lemma 5.5. *The step sizes can be chosen to satisfy $\tilde{z}(y_{[r]}) = 0$ eventually. Hence eventually $\pi(y_{[r]}) = \pi'$ is constant.*

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

Lemma 5.6. *Suppose Assumption 5.1 holds, and let (q, u) be a cluster point of $\{(y_{[r]}, v_{[r]})\}$. Then $q \in P_\mathcal{D}$, if $h(\cdot, u; q) \geq 0$ on $Z(q)$.*

Proof. Since $\{f_\nu(y_{[r]})\}$ is bounded from below by assumption, and monotonically decreasing by Lemma 5.2, it holds that

$$\lim_{r\to\infty} (f_\nu(y_{[r]}) - f_\nu(y'_{[r]})) = 0. \tag{5.12}$$

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

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 $(y_{[r_\ell]}, v_{[r_\ell]}) \to (q, u)$, we get $T_\omega(y_{[r_\ell]}, v_{[r_\ell]}) \to T_\omega(q, u)$. Therefore $\lim f_\nu(y'_{[r_\ell]}) = f_\nu(q') < f_\nu(q) = \lim f_\nu(y_{[r_\ell]})$, which contradicts (5.12). 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 $[T_1(y_{[r_\ell]}, v_{[r_\ell]}), T_{\overline{\omega}}(y_{[r_\ell]}, v_{[r_\ell]})] \ni T_{\omega_{r_\ell}}(y_{[r_\ell]}, v_{[r_\ell]}) = y'_{[r_\ell]}$ 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 $\rho_{\pi(q)}\bar{\rho}_{\pi'} z(y_{[r_\ell]}) = -\rho_{\pi(q)}\bar{\rho}_{\pi'} S_\pi^\dagger(y_{[r_\ell]}) g_\pi^{v[r_\ell]}(y_{[r_\ell]}) \to 0$, because $v_{[r_\ell]} \to u$ is bounded, and $\rho_{\pi(q)}\bar{\rho}_{\pi'} S_\pi^\dagger(y_{[r_\ell]})$ goes to zero (with $1/d_k(y_{[r_\ell]})$ going to infinity in $S_\pi(y_{[r_\ell]})$ for $k \in \pi(q) \setminus \pi'$). As $\bar{\rho}_{\pi(q)} S_\pi^\dagger(y_{[r_\ell]}) g_\pi^{v[r_\ell]}(y_{[r_\ell]})$ does not depend on a_k for $k \in \pi(q)$, it is convergent. Therefore, in summary, we have $\bar{\rho}_{\pi'} S_\pi^\dagger(y_{[r_\ell]}) g_\pi^{v[r_\ell]}(y_{[r_\ell]}) \to \bar{\rho}_{\pi(q)} S_\pi^\dagger(q) g_\pi^u(q)$.

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

Lemma 5.7. *Suppose Assumption 5.1 holds. Let Q_k denote the set of cluster points (q, u) of $\{(y_{[r]}, v_{[r]})\}$, such that $k \in \pi(q) \setminus \pi'$. We have,*

(i) If $\liminf_{\ell\to\infty} d_k(y'_{[r_\ell]})/d_k(y_{[r_\ell]}) > 1$ 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 $\tilde{z} \in Z_k(q) \triangleq \{\tilde{z} \in Z(q) \mid W_i\tilde{z} = 0$ 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 $\{(y_{[r]}, v_{[r]})\}$, and that $Z_k(q) = \{\tilde{z} \in Z(q) \mid \rho_{\pi'}\tilde{z} = 0, \pi_{\tilde{z}} = \{k\}\}$. Let $\{(y_{[r_\ell]}, v_{[r_\ell]})\}$ be a subsequence of $\{(y_{[r]}, v_{[r]})\}$ approaching Q_k (with constant $\pi(y_{[r_\ell]}) = \pi'$). Under the conditions of (ii), we must have $\liminf_{\ell\to\infty} d_k(y'_{[r_\ell]})/d_k(y_{[r_\ell]}) > 1$, because otherwise we could find a subsequence convergent to some $(q, u) \in Q_k$, for which an application of Lemma 5.3 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 $(y_{[r]}, v_{[r]}) \in Q_k + \mathbb{B}(0, \delta)$, then $d_k(y'_{[r]}) \geq (1 + \epsilon)d_k(y_{[r]})$. Therefore, since $d_k(y_{[r]}) > 0$, there exists a $t > r$ such that $(y_{[t]}, v_{[t]}) \notin Q_k + \mathbb{B}(0, \delta)$. Thus the whole sequence cannot converge to Q_k.

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

$$f_v(y_{[r]}) - f_v(y'_{[r]}) \geq f_{\tilde{v}[r]}(y_{[r]}) - f_{\tilde{v}[r]}(y'_{[r]}) = \frac{1}{2} \sum_{i \notin \pi(y_{[r]})} (d_i(y_{[r]}) - d_i(y'_{[r]}))^2 / d_i(y_{[r]}), \tag{5.13}$$

where the final estimate and term $C/2$ are from the proof of Theorem 5.1. This with $r = r_\ell$ provides a contradiction to (5.12). Therefore $Q_k = \emptyset$. □

Theorem 5.2. *Suppose Assumption 5.1 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 $\{(y_{[r]}, v_{[r]})\}$ 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 5.6 proves the claim.

Otherwise, to reach a contradiction, we may assume that $(y_{[r]}, v_{[r]}) \to (q, u)$, where $h(\tilde{z}, u; q) < 0$ for some $\tilde{z} \in Z(q)$. According to Lemma 5.4, 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

[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 (5.7). The argument of Lemma 5.7 could also be extended to allow $k \in \pi(y_{[r]})$, provided $\|\alpha\tilde{z}\|_k > 0$ for a subsequence. However, application/variant of Lemma 5.4 would demand additional assumptions.

$\tilde{z} \in Z_k(q)$ for some $k \in \pi(q) \setminus \pi'$. An application of Lemma 5.7 to $Q_k = \{(q, u)\}$ now provides the desired contradiction. □

Remark 5.1. We have the following further observations:

(i) If $y_{[r]} \to q$, 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 5.6). In particular, any cluster point with $\pi(q) = \emptyset$, is a solution.

(iii) If $\#\pi(y) \leq 1$ for all $y \in \mathbb{R}^m$, then there is a cluster point $q \in P_{\mathcal{D}}$. (Combine Lemmas 5.6 and 5.7.) 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 $d_k(y_{[r]}) > \delta$, as well as $d_k(y'_{[r_\ell]}) \in [\delta/2, \delta)$, since $d_k(y'_{[r_\ell]}) \to 0$ is not possible by (5.13). Therefore there are cluster points in this distance range. Now let $\delta \searrow 0$.

5.6 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 $\{f_\nu(y_{[r]})\}_{r=1}^{\infty}$ is descending, it suffices to show that the level sets of f_ν are bounded. This is where we need the general results of Section 2.5, 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 (5.3).

Lemma 5.8. *Let $A \triangleq \bigcup_{y \in \mathbb{R}^m} \partial f(y)$. Then* 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

$$Z_{\pi_b} = \{ \sum_{k \notin \pi_b} W_k^2 q / \|q\|_k + v \mid q \in Q_{\pi_b}, v \in \operatorname{cl} A_{\pi_b} \},$$
$$Q_{\pi_b} \triangleq \{q \in \mathbb{R}^m \mid W_j q = 0 \ (j \in \pi_b), W_k q \neq 0 \ (k \notin \pi_b)\},$$
$$A_{\pi_b} \triangleq \bigcup_{y \in \mathbb{R}^m} \partial(\sum_{k \in \pi_b} d_k)(y).$$

Proof. The subdifferentials of f are clearly uniformly bounded: for $g \in \partial f(y)$, $\|g\| \leq \sum_{k=1}^n \|W_k\|$. Hence A is bounded. By, e.g., [Rockafellar, 1972, Section 24] cl A is also convex. It remains to prove that bd A is of the claimed form.

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

$$g = \sum_{k \notin \pi_b} W_k^2 q / \|q\|_k + v \tag{5.14}$$

with $\pi_b = \{j \in \{1, \ldots, n\} \mid W_j q = 0\}$, and $v \in \mathrm{cl}\, A_{\pi_b}$. Therefore, all the exposed faces of $\mathrm{cl}\, A$ 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 $\mathrm{bd}\, A$.

The exposed faces of $\mathrm{cl}\, A$ are precisely the sets of the form $\mathrm{cl}\, A \cap H$, where H is a supporting hyperplane to $\mathrm{cl}\, A$ [see Rockafellar, 1972]. But $\mathrm{cl}\, A$ is the intersection of the corresponding half-spaces. Thus, if $g \in \mathrm{cl}\, A$ has a ball $\mathbb{B}(g, \epsilon)$ around it that is not intersected by any of the hyperplanes H (and thus not by any of the Z_{π_b}), then $g \notin \mathrm{bd}\, A$. Otherwise, since the intersecting hyperplanes are defined by a compact set of parameters (closed subset of $\mathrm{bd}\, A \times \mathrm{bd}\, \mathbb{B}(0, 1)$), we may find a supporting hyperplane H that contains g. But then $g \in \mathrm{cl}\, A \cap H$, an exposed face. □

6 CLUSTERING APPLICATIONS

6.1 Introduction

As already discussed in Chapters 1 and 5, the general theme of the present chapter is the problem of locating one or more points $y_1, \ldots, y_K$ according to some optimality criterion involving another set of n fixed points and combinations of distances between them. Furthermore, we require that the problem can be modelled in the form (5.1), to study the application of the perturbed Weiszfeld method.

In the single-prototype case ($K = 1$), popular objectives are the data means and the 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 [Weiszfeld, 1937; Kuhn, 1973]. Multi-prototype ($K > 1$) variants of the location problem often somehow involve the single-facility case. In particular, in case of criteria of the *K*-means type [Cox, 1957; Selim and Ismail, 1984], the goal is to assign each vertex to the closest prototype y_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 Teboulle [2007].

The classic multisource Weber problem – the problem of finding the *K-spatial-medians* – otherwise also known as the location-allocation problem [Cooper, 1964], is a problem of *K*-means type. The distance between a point and a prototype is merely taken to be the Euclidean instead of squared distance. Indeed, the problem is of the form (5.1), and in our analysis of the problem in Section 6.3, it furthermore turns out that the perturbed Weiszfeld method in this case almost reduces into a single-step variant of the *K*-means style algorithm if the Weiszfeld method were to be applied on each cluster.

We also analyse in this chapter – the following Section 6.2 more specifically – a new clustering problem of the form (5.1). The objective is based on a multi-objective approach to the general problem: a mathematical statement of "place prototypes close to data and far from each other". After the analysis of these two

clustering formulations, we finish the chapter with a few experimental comparisons presented in Section 6.4.

6.2 Bi-objective clustering

Consider a multiobjective formulation of the multifacility location problem:

$$\min_{\bar{y}\in(\mathbb{R}^m)^K} (f_1, f_2)(\bar{y}; \bar{a}), \tag{6.1}$$

where the minimum is in the sense of Pareto-optimality, $\bar{y} = (y_1, \ldots, y_K) \in (\mathbb{R}^m)^K$, and $\bar{a} = (a_1, \ldots, a_n) \in (\mathbb{R}^m)^n$. The objectives are defined as

$$f_1(\bar{y}) \triangleq \sum_{i=1}^{K} \sum_{j=1}^{n} d_j(y_i), \qquad f_2(\bar{y}) \triangleq -\frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} d(y_j, y_i)$$

for some distance functions d and d_j, the latter dependent on a_j. The objective f_1 indicates our desire to place cluster centres $\{y_j\}$ 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 $-f_2$.)

6.2.1 Squared Euclidean distance

Although it does not fit in the framework of the perturbed spatial median, for comparison to what will follow and also to the classical K-means, 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., Miettinen, 1999, Chapter I.3] that

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

or that

$$(\lambda_1 n - \lambda_2 K) y_i - \lambda_1 \sum_{j=1}^{n} a_j + \lambda_2 \sum_{j=1}^{K} y_j = 0 \quad \text{for all } i = 1, \ldots, K, \tag{6.2}$$

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

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

$$\frac{1}{K} \sum_{i=1}^{K} y_i = \frac{1}{n} \sum_{j=1}^{n} a_j. \tag{6.3}$$

If, on the other hand $\lambda_1 n - \lambda_2 K \neq 0$, we find that all the $\{y_i\}_{i=1}^K$ are equal by subtracting the term (6.2) for y_i and y_j $(i \neq j)$. Hence, unless $\lambda_1 = 0$, in fact (6.3) holds. In the case $\lambda_1 = 0$ there is no finite minimum, so we may ignore it.

Let us now check when solutions of (6.3) are Pareto-optimal. Expand the expressions for d to yield

$$f_1(\bar{y}) = \sum_{i=1}^{K}\sum_{j=1}^{n}\frac{1}{2}(\|y_i\|^2 + \|a_j\|^2) - (\sum_{i=1}^{K} y_i)^T(\sum_{j=1}^{n} a_j)$$

and

$$2f_2(\bar{y}) = -\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{1}{2}(\|y_i\|^2 + \|y_j\|^2) + (\sum_{i=1}^{K} y_i)^T(\sum_{j=1}^{K} y_j).$$

Thus if (6.3) holds, then both f_1 and f_2 have a constant term at the end and f_2 decreases if and only if $\sum_i \|y_i\|^2$ increases. But since this means that f_1 increases, the solutions of (6.3) 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.

6.2.2 Euclidean distance

With the Euclidean distance $d(x, y) = \|x - y\|$, and d_j defined as in Section 5.2, we get more interesting results. After rewriting $\nu_{\mathrm{MO}} \triangleq -f_2$, the scalarisation of the problem (6.1) by the factor $\lambda \geq 0$ then reads as [cf. Miettinen, 1999, Section II.3.1]

$$\min f_1(\bar{y}) - \lambda\nu_{\mathrm{MO}}(\bar{y}). \tag{6.4}$$

This problem can be cast as a problem of finding a perturbed spatial median as follows. For each $i = 1, \ldots, K$ and $j = 1, \ldots, n$, let

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

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

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

Because ν_{MO} is convex and finite, the problem can be modelled as a perturbed spatial median problem with vertices $\{a_j^i\}$ and perturbation $\lambda\nu_{\mathrm{MO}}$. Note that if $\{W_j\}$ satisfy the range non-overlap assumption of Theorem 5.2, so do $\{W_j^i\}$. Hence, by Theorem 5.2, Theorem 2.6, and Lemma 2.9, Algorithm 5.1 is applicable for finding semi-critical points (Kuhn-Tucker points of the multiobjective problem), if we can bound $\mathcal{R}(\partial(\lambda\nu_{\mathrm{MO}}))$ within $\mathcal{R}(\partial f_1)$.

With f denoting here and throughout the chapter, the function defined by (5.3) with the original data $\{a_j\}$, not $\{a_j^i\}$, we note that $\partial f_1(\bar{y}) = \partial f(y_1) \times \cdots \times$

$\partial f(y_K)$, since f_1 consists of K sets of n terms depending on different components of $\bar{y}$. Also, since ν_{MO} is positively homogeneous, we have $\mathcal{R}(\partial(\lambda\nu_{\mathrm{MO}})) = \partial(\lambda\nu_{\mathrm{MO}})(0)$. Now, since $\mathcal{R}(\partial f_1)$ is a product space, it suffices to consider the slices $[\partial(\lambda\nu_{\mathrm{MO}})(0)]_i$ of this subdifferential independently. At differentiable points

$$[\nabla(\lambda\nu_{\mathrm{MO}})(\bar{y})]_i = \lambda \sum_{j\neq i} \frac{y_i - y_j}{\|y_i - y_j\|}.$$

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

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

or that

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

In the simple case with $W_k = I$ for all $k = 1, \dots, n$, this follows if $\lambda < n/(K-1)$ (when $K > 1$), because $\operatorname{cl} \mathcal{R}(\partial f) = \mathbb{B}(0, n)$ then. For incomplete and weighted data, we must consider the "minimal dimension" of A: by Lemma 5.8, 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 6.1.

(i) The level sets of the scalarised problem (6.4) *are bounded if* $0 \leq \lambda < \beta/(K-1)$ *with* $\beta = \min \|\sum_{k\notin\pi_b} W_k^2 q/\|q\|_k\|$, *with the minimum taken over all* $q \in Q_{\pi_b}$ *and* $\pi_b \subset \{1, \dots, n\}$ *satisfying the conditions of Lemma 5.8.*

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

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

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 5.8, 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 5.1 can thus in principle be applied to finding semi-critical points of the scalarised problem (6.4). We emphasise that Theorem 6.1(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 $\lambda > \beta/(K-1)$, $\lambda \sum_{j=1}^{K-1} q_j/\|q_j\| \in \operatorname{cl} \mathcal{R}(\partial f)$ can be violated, whence $\mathcal{R}(\partial(\lambda \nu_{\mathrm{MO}})) \not\subset \mathcal{R}(\partial f_1)$. Problem (6.4) is not bounded from below then, wherefore no finite Pareto-optimal solution is generated by scalarisation parameters much larger than Algorithm 5.1 can be expected to handle.

Remark 6.1. 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 $g_\pi(\bar{y}) = (g_\pi(y_1), \ldots, g_\pi(y_K))$ and $S_\pi(\bar{y}) = (S_\pi(y_1), \ldots, S_\pi(y_K))$, where the right-hand-sides have been defined for the original data $\{a_i\}$. In consequence, there is no dependency between the y_j within the iterations of the SOR-Weiszfeld algorithm aside from calculating the "tilt" $v \in \partial \nu_{\mathrm{MO}}(\bar{y})$. Therefore each iteration of Algorithm 5.1 can be calculated in parallel using the same step size for the different cluster centres after a subgradient of ν_{MO} has been calculated.

Remark 6.2. 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 6.1. *Suppose $y_j = y_k$ ($j \neq k$) and $\operatorname{rank}(\rho_\pi) < m$ for $\pi \triangleq \pi(y_j) = \pi(y_k)$. Then $\bar{y}$ is not a local minimiser.*

Proof. The term $\|y_j - y_k\|$ is not differentiable at $y_j = y_k$, Therefore, with $\bar{v} = (v_1, \ldots, v_K)$, there are multiple choices for v_j and v_k (dependent on each other) in all m dimensions, and we can in (5.5) choose v_j so that $[g_\pi(\bar{y})]_j - v_j \neq 0$, 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 (5.5). Thus the claim of the lemma follows from Theorem 5.1. □

6.3 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

$$\min_{w_{ij}, \bar{y}} \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij} d_i(y_j) \quad \text{with } w_{ij} \in \{0, 1\} \text{ and } \sum_{j=1}^{K} w_{ij} = 1. \tag{6.5}$$

The weights w_{ij} indicate to which cluster j the vertex i belongs to, and y_j is the cluster prototype.

The standard K-means-type algorithm [Cox, 1957; Selim and Ismail, 1984; Cooper, 1964] proceeds by assigning each a_i to the closest cluster centre y_j (setting $w_{ij} = 1$), calculating the spatial median y'_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 Selim and Ismail [1984], along with providing an extension to find local minima. The proof readily generalises to our case of incomplete data (but see also Appendix 1). For some other heuristic and local methods for solving the problem, see Cooper [1964]; Brimberg et al. [2000]; Bongartz et al. [1994]. The global solution with outer approximation methods of the diff-convex formulation to be given below is studied in Chen et al. [1998]. Other approximation schemes are derived in Arora et al. [1998].

Given the constraints on the weights, for fixed i, $\min_{w_{ij}} \sum_{j=1}^{K} w_{ij} d_i(y_j) = \min_{j=1,\ldots,K} d_i(y_j)$. Therefore an alternative way to write (6.5) is

$$\min_{\bar{y}} \sum_{i=1}^{n} \min_{j=1,\ldots,K} d_i(y_j). \tag{6.6}$$

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

$$f_1(\bar{y}) - \nu_{\mathrm{KM}}(\bar{y}) \triangleq \Big(\sum_{i=1}^{n} \sum_{j=1}^{K} d_i(y_j)\Big) - \Big(\sum_{i=1}^{n} \max_{j=1,\ldots,K} \sum_{k \neq j} d_i(y_k)\Big).$$

But, indeed, using the lifting of a_i to a_i^j for $j = 1, \ldots, K$ as in Section 6.2.2, this problem is seen to be a problem of perturbed spatial median. This problem, however, has unbounded level sets: any change in y_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 Brimberg and Mladenović [1999]. Therefore Theorem 2.6 cannot 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 $\bar{v}(\bar{y})$, after first analysing Algorithm 5.1 applied to this problem, in further detail.

6.3.1 Algorithm analysis and reduction

Let us calculate $\partial \nu_{\mathrm{KM}}$. Similarly to the derivation of ∂f_1 in Section 6.2.2, we get

$$\partial\Big(\sum_{k \neq j} d_i(y_k)\Big)(\bar{y}) = \partial d_i(y_1) \times \cdots \partial d_i(y_{j-1}) \times \{0\} \times \partial d_i(y_{j+1}) \times \cdots \partial d_i(y_K)$$

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

$$\begin{aligned}\partial \nu_{\mathrm{KM}}(\bar{y}) &= \bigcup_{\Lambda\in\mathcal{W}} \sum_{i=1}^{n} \sum_{j\in J_i} \lambda_{j,i} \partial\big(\sum_{k\neq j} d_i(y_k)\big)(\bar{y}) \\ &= \bigcup_{\Lambda\in\mathcal{W}} \sum_{i=1}^{n} \prod_{j=1}^{K} \left(\sum_{k\in J_i\setminus\{j\}} \lambda_{k,i} \right) \partial d_i(y_j) = \bigcup_{\Lambda\in\mathcal{W}} \sum_{i=1}^{n} \prod_{j=1}^{K} G_{j,i}\end{aligned} \tag{6.7}$$

with

$$G_{j,i} = \begin{cases} \partial d_i(y_j), & j \notin J_i, \\ (1-\lambda_{j,i})\partial d_i(y_j), & j \in J_i. \end{cases} \tag{6.8}$$

Here $\Lambda \triangleq \{\lambda_{j,i} \mid j \in J_i, i = 1,\dots,n\}$ and $\mathcal{W} \triangleq \mathcal{W}(\bar{y}) \triangleq \{\Lambda \mid \sum_{j\in J_i} \lambda_{j,i} = 1, \lambda_{j,i} \geq 0\}$. 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 $\bar{v}(\bar{y}) = \bar{v} = (v_1,\dots,v_K)$ each $v_j \in \sum_{i=1}^{n} G_{j,i}$ independently. Noting that $j \notin J_i$ implies $d_i(y_j) > 0$ and hence $i \notin \pi(y_j)$, let

$$v_j \triangleq \sum_{\substack{i\notin\pi(y_j)\\ J_i\ni j}} (1-\lambda_{j,i})\nabla d_i(y_j) + \sum_{\substack{i\notin\pi(y_j)\\ J_i\not\ni j}} \nabla d_i(y_j) + \sum_{i\in\pi(y_j)} (1-\lambda_{j,i}) W_i^2 z_j / \|z_j\|_i. \tag{6.9}$$

Then $\bar{v} \in \partial\nu_{\mathrm{KM}}(\bar{y})$, and $\bar{\rho}_{\pi(\bar{y})} g_\pi^v(\bar{y}) = (\bar{\rho}_{\pi(y_1)} g_1, \dots, \bar{\rho}_{\pi(y_K)} g_K)$ for

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

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

$$h(z_j, v_j; y_j) = \Big(\sum_{i\notin\pi(y_j)} \nabla d_i(y_j) - v_j\Big)^T z_j + \sum_{i\in\pi(y_j)} \|z_j\|_i = g_j^T z_j + \sum_{\substack{i\in\pi(y_j)\\ J_i\ni j}} \lambda_{j,i} \|z_j\|_i, \tag{6.10}$$

which are h for the same reduced problems. It follows that $\tilde{z}$ required by Lemma 5.1 can be chosen independently for each j, together with v_j. (Note that v_j only depends on the $\mathcal{R}(\rho_{\pi(y_j)})$ part of z_j, i.e., $\tilde{z}_j$.) However, $S_\pi(\bar{y})$ does not split into clusters quite so well: it remains dependent on the whole original data set, $S_\pi(\bar{y}) = (S_{\pi,\mathrm{full}}(y_1),\dots,S_{\pi,\mathrm{full}}(y_j))$, where $S_{\pi,\mathrm{full}}(y_j) \triangleq \sum_{i\in\{1,\dots,n\}\setminus\pi(y_j)} S_i(y_K)$. Despite this, the direction of (5.6),

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

can be calculated almost independently for each j. We have therefore showed

Theorem 6.2. *For the multisource Weber problem, Algorithm 5.1 reduces to calculating at each step, for the spatial median problems*

$$\min_{y'_j} \sum_{i:j\in J_i} \lambda_{j,i} d_i(y'_j) \quad (j = 1, \ldots, K), \tag{6.11}$$

one iteration starting from y_j, of the convex SOR-Weiszfeld algorithm, modified to use $S_{\pi,\text{full}}(y_j)$ (instead of $S_\pi(y_j)$ for the data set A_j), in the direction of (5.6)*. 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 (5.6) and (5.7) (for the problem (6.11)), 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 problem (6.11).

6.3.2 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. Kuhn [1973]. 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 $\pi(y) = \emptyset$ and convex problems of spatial medians, each coordinate of $T_1(y)$ 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 $\hat{y} \in \mathbb{R}^m$ be some reference point, e.g., a spatial median of the data, $L > 1$, and $\pi \triangleq \pi(y_j)$. Then, for the difference of $y'_j \triangleq T_\omega(y_j)$ and $\hat{y}$, following (5.10), we have for the coordinates k present in $\mathcal{R}(\bar{\rho}_\pi)$ that

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

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

To bound $(y'_j - \hat{y})_k$ for coordinates in $\mathcal{R}(\rho_\pi)$, we alter the parameter α in the iteration. By the definition of the step $z(y)$ in (5.6), $|(y'_j - \hat{y})_k| \leq |(y_j - \hat{y})_k| + \alpha\omega$.

By Lemma 5.1, 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 $|(y_j - \hat{y})_k| < LC_k - \alpha\omega$, we have $|(y'_j - \hat{y})_k| < LC_k$. Hence we can state:

Theorem 6.3. *With the choice of α and ω as above, the sequence of iterates for the perturbed SOR-Weiszfeld algorithm of Theorem 6.2 can be held bounded for the K-spatial-medians objective. In consequence, the convergence results of Theorem 5.2 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 $(y - \hat{y}) + \prod_{k=1}^{m}(-LC_k, LC_k)$ 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 is enough leeway for ω for it to be altered in such a manner that the conditions in Theorem 5.2 on ω are met. Furthermore, the K-spatial-medians objective function clearly is bounded from below, so the theorem applies.

In the choice (6.9) of v_j used to obtain (6.10), we choose $W_i^2\tilde{z}/\|\tilde{z}\|_i \in \partial d_i(y_j)$ for $i \in \pi(y_j)$. These are in the limit of gradients of differentiable points of ν_{KM}, for at these points $\nabla d_i(y_j)$ takes the form $W_i^2(y_j - a_i)/d_i(y_j)$. Furthermore, directions in $\mathcal{D}_{\text{N}}(-f_{\text{KM}}^2)(\bar{y})$ have $\Lambda \in \mathcal{W}_{\text{ext}}$. For, if $\#J_i(\bar{y}) > 1$, then ν_{KM} is not differentiable, and hence at differentiable points $\mathcal{W} = \mathcal{W}_{\text{ext}}$. As directions in $\mathcal{D}_{\text{N}}$ are limits of directions at differentiable points, by the preceding we must have have $\Lambda \in \mathcal{W}_{\text{ext}}$ for such directions. Now, if $\bar{y}$ is $\mathcal{D}_{\text{N}}$-critical, then there is a choice of weights $\Lambda \in \mathcal{W}_{\text{ext}}$ for which $(\bar{y}, \Lambda)$ solves (6.11) for each j. Therefore, with such choice of Λ, $\bar{v} \in \mathcal{D}_N(\bar{y})$. □

6.3.3 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 6.2 in a combined form as finding $\bar{y}'$ such that $F(\bar{y}';\Lambda) < F(\bar{y};\Lambda)$ for the function

$$F(\bar{y};\Lambda) \triangleq \sum_j \sum_i \lambda_{j,i} d_i(y_j). \tag{6.12}$$

Theorem 6.4. *The point $\bar{y}^*$ is a local minimum of* (6.6) *if and only if it minimises $F(\cdot;\Lambda)$ for all $\Lambda \in \mathcal{W}(\bar{y}^*)$.*

Proof. Necessity is obvious: $(f_1 - \nu_{\text{KM}})(\bar{y}) = \sum_{i=1}^{K} \min_j d_i(y_j) \leq F(\bar{y};\Lambda)$ with equality at $\bar{y}^*$, for all $\Lambda \in \mathcal{W}(\bar{y}^*)$. Hence if $\bar{y}^*$ is not a minimiser of the convex function $F(\cdot;\Lambda)$ for some such Λ, it cannot minimise (6.6) even locally.

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

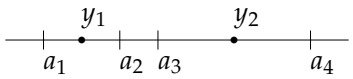
Corollary 6.1. *(i) If* $\#J_i(\bar{y}^*) = 1$ *for all* $i = 1, \ldots, n$*, and* $\bar{y}^*$ *minimises* $F(\cdot; \Lambda^*)$ *for the unique* $\Lambda^* \in \mathcal{W}(\bar{y}^*)$*, then* $\bar{y}^*$ *is a local minimiser of* (6.6)*. (ii) If* $\#J_i(\bar{y}) > 1$*, and we have* $\pi(y_j) = \emptyset$ *for some* $j \in J_i(\bar{y})$*, then* $\bar{y}$ *is not a local minimiser.*

Proof. The first claim is obvious from the preceding theorem. As for the second claim, suppose $\bar{y}$ minimises $F(\cdot; \Lambda^*)$ for some $\Lambda^* \in \mathcal{W}(\bar{y})$. Let $j, j' \in J_i(\bar{y})$, $j \neq j'$, and $\pi(y_j) = \emptyset$. 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 $\bar{y}$. However, the condition $0 \in \{\nabla \sum_i \lambda_{j,i} d(a_i, y_j)\}$ will be upset, and hence the value of $f_1 - \nu_{\mathrm{KM}}$ can be improved locally. □

Corollary 6.2. *If* $\bar{y}^*$ *is* $\mathcal{D}_{\mathrm{N}}$*-critical and* $\#J_i(\bar{y}^*) = 1$ *for all* i*, then* $\bar{y}^*$ *is a local solution of* (6.6)*.*

Proof. The condition $\#J_i(\bar{y}^*) = 1$ forces Λ^* to be unique. Therefore $(1 - \lambda_{j,i})\partial d_i(y_j^*)$ also reduces to the singleton $\{0\}$ in (6.8). Hence $\bar{v}(\bar{y}^*)$ is uniquely determined. It then follows from $\mathcal{D}_N$-criticality that $\bar{y}^*$ minimises (6.11) for all j, and consequently minimises (6.12). That $\bar{y}^*$ is a local solution follows from Corollary 6.1. □

Remark 6.3. In fact, that $\#J_i(y^*) = 1$ or $\min_j d_i(y_j^*) > 0$ for all i forces $\bar{v}(\bar{y}^*)$ 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 y_1 and y_2 are at equal distance from a_3. If a_3 is assigned to the cluster of y_2, we have a critical point, yet assignment to y_1 shows that both cluster centres can be improved by just a small move of either or both y_1 or y_2 to the right.

Corollary 6.3. *Under conditions of Theorem 6.3, with choices of* $\Lambda \in \mathcal{W}_{\mathrm{ext}}(\bar{y})$*, if the iterates* $\{\bar{y}_{[r]}\}$ *of the algorithm of Theorem 6.2 converge to* $\bar{y}^*$*, then it is either a local minimiser, or has disputed vertices:* $\#J_i(\bar{y}^*) > 1$ *for some* $i \in \{1, \ldots, n\}$*.*

Proof. Since $\{\bar{y}_{[r]}\}$ converge to $\bar{y}^*$, if $\{\bar{v}(\bar{y}_{[r]})\}$ diverges, then ν_{KM} is non-differentiable at $\bar{y}^*$ (cf. Remark 5.1(i)). This says that there are disputed vertices. If $\{\bar{v}(\bar{y}_{[r]})\}$ also converges, then by Theorem 6.3 (and Theorem 5.2), $\bar{y}^*$ is $\mathcal{D}_{\mathrm{N}}$-critical, and the previous corollary applies. □

Remark 6.4. Suppose that eventually in the method, the assignments of vertices to clusters is unique. Then, if the data set is complete (or more generally $\#\pi(y_i) \leq 1$ always), we have convergence to the set of local minimisers (being able to analyse the method on each cluster separately, applying Remark 5.1(iii)). Therefore, with such simple data, non-convergence is always a case of dispute over assignment of vertices to clusters.

6.3.4 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 K 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 K "tilted" SOR-Weiszfeld iterations in parallel, with tilts calculated from the results of all the K previous iterations, as was explained in Section 6.2.2.

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 5.1 reduces to the following:

Algorithm 6.1 (K-means type method with single step SOR-Weiszfeld)**.**

1. Choose some starting points y_j ($j = 1, \ldots, K$).

2. Assign each vertex a_i ($i = 1, \ldots, n$) to one of the clusters A_j corresponding to closest y_j ($j = 1, \ldots, K$).

3. To obtain y'_j, calculate for the (convex) spatial median problem on A_j, one iteration of Algorithm 5.1 with the modified direction

$$z^{\mathrm{KM}}(y_j) \triangleq -\bar{\rho}_{\pi(y_j)} S^{\dagger}_{\pi,\mathrm{full}}(y_j) g_\pi(y_j) + \alpha \rho_{\pi(y_j)} \tilde{z}_j, \tag{6.13}$$

   where $g_\pi(y_j)$ 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 6.2. Since (5.7) defining the bound α_2 for the whole problem is the sum of $S_{\pi,\mathrm{full}}$-modified conditions for the sub-problems, it suffices to bound α from above by the minimum of the upper bounds for the sub-problems. Theorem 5.2 sets some minor restrictions on $\omega \in [1, 2)$ to avoid oscillation. Theorem 6.3 sets additional upper bounds on the step lengths by the coordinate-wise bound $LC_k > C_k$ on $|(\dot{y}'_j)_k|$, which we may, however, choose arbitrarily large.

Example 6.1. When $W_k = w_k I$ for all $k = 1, \ldots, n$, $S_{\pi,\mathrm{full}}$ is proportional to the identity; cf. Example 5.1. Therefore, in that case, (6.13) 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 6.2, it is interesting to interpret the K-spatial-medians as one scalarisation of a more general problem

$$\min_{\bar{y}}(f_1, -\nu_{\mathrm{KM}})(\bar{y}).$$

The meaning of the objective f_1 is the same as before. What the objective $-\nu_{\mathrm{KM}}$ 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 $\min_{\bar{y}} f_1(\bar{y}) - \lambda\nu_{\mathrm{KM}}(\bar{y})$ for $\lambda \in [0, 1]$.

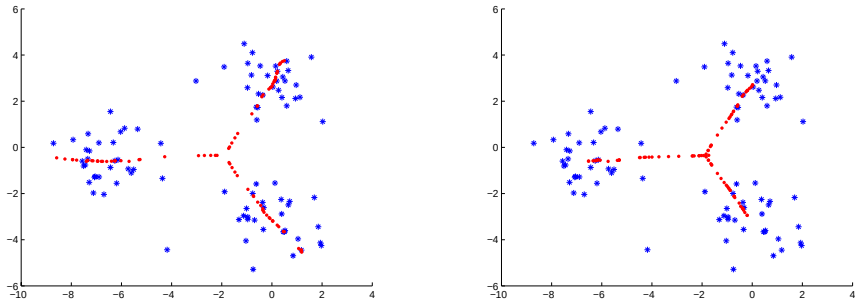
For $\lambda \in [0, 1)$, Theorem 2.6 is applicable to proving boundedness of the level sets. To see this, consider the inclusions $\lambda\mathcal{R}(\partial\nu_{\mathrm{KM}}) \subset \lambda \bigcup_\Lambda \mathcal{R}(\partial f_\Lambda) \subset \operatorname{int}\mathcal{R}(\partial f_1)$, where $f_\Lambda : \bar{y} \mapsto \sum_{j=1}^{K}\sum_{i=1}^{n}(1 - \lambda_{j,i})d_i(y_j)$, 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 $\bar{y}$ in the expression (6.7). 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 2.6 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., $\operatorname{cl}\mathcal{R}(\partial(\lambda\nu_{\mathrm{KM}})) \subset \operatorname{int}\mathcal{R}(\partial f_1)$. Now Theorem 2.6 applies again.

6.4 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 Äyrämö [2006].

Figures 6.1 and 6.2 show results for two cases using both the problem of Section 6.2.2 (MO), and the multi-objective formulation of the K-spatial-medians (KM) discussed in Section 6.3. 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 $\max_{j=1,2,3} \|y'_j - y_j\| < 10^{-6}$ and the maximum number of iterations was 300. The 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.

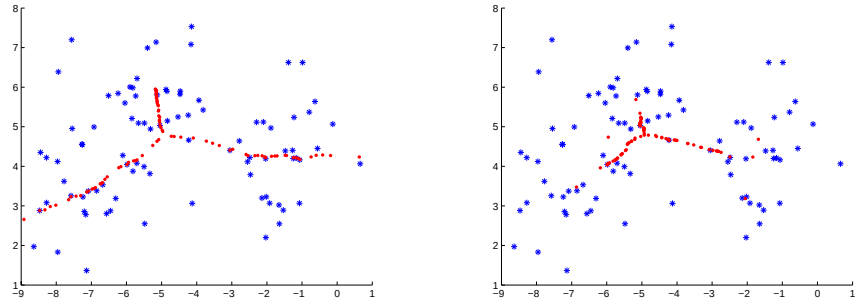
As we can see, for $\lambda = 0$ the result is in both cases the spatial median of the data. From there, the solutions continuously move towards the centres of clusters, as λ varies towards the respective upper bound for λ ($\lambda_{\mathrm{sup}} = n/(K-1)$ for MO, from Theorem 6.1, and $\lambda_{\max} = 1$ for KM, from the analysis of Section 6.3.2), just



nits: 57/42/27/191
(a) MO, $\lambda_{\text{sup}} = 90/2$

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

FIGURE 6.1 Results for a task with three clear clusters with varying λ and $\omega = 1.5$.



nits: 68.8/63/39/151
(a) MO, $\lambda_{\text{sup}} = 90/2$

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

FIGURE 6.2 Results for a task with three less clear clusters with varying λ and $\omega = 1.5$.

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; see Section 2.4 for the subdifferential and epigraphical approach, and, e.g., Bonnans and Shapiro [1998] for a review of other results. 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 is not entirely unexpected, however: for tightly packed clusters, we should have $d(y_k^*, y_j^*) \approx \sum_{i:k\in J_i} d(a_i, y_j^*)/\#\{i : k \in J_i\}$ 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 [Kärkkäinen and Äyrämö, 2005]. 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.

7 THE EUCLIDEAN TRAVELLING SALESPERSON PROBLEM

7.1 Introduction

This chapter is concerned with the travelling salesperson problem with Euclidean (ℓ^2) distances (undiscretised), i.e., the problem of finding the shortest closed path that visits every vertex (or city) in a given finite subset of $\mathbb{R}^m$ exactly once, with the distances given by the Euclidean metric. Whereas various rather efficient algorithms exist for the general and general metric TSP [Johnson and McGeoch, 2002], few seem to be able to take advantage of the special features of the variant with Euclidean distances – that still remains NP-hard. The most remarkable of those that do are Arora's [1998; 2003], polynomial time (and even "nearly linear time") approximation schemes (PTAS) the good performance of which is, however, only asymptotic. Other methods for Euclidean instances specifically include various heuristics optimised for speed and based on clustering or partitioning of the plane, or spacefilling curves.

Here, we make another stab at formulating and finding (local) solutions to the Euclidean TSP. Our approach consists of first reformulating the problem as a continuous diff-convex problem. Instead of attempting to find the optimal path, we attempt to find points that construct the path, constrained to equal one of the input vertices. We then relax this problem, converting the constraint into a mere penalty. Dependent on the formulation of the constraint, the relaxed problem is found to be equivalent to certain clustering problems (including the multisource Weber problem or "K-spatial medians") perturbed with the path length penalty. (Perhaps not so coincidentally, Arora's methods can also be extended to approximate the K-spatial medians [Arora et al., 1998; Arora, 2003].)

As a continuation of the work in Chapters 5 and 6, in the present chapter

[0] This chapter is based on the article Valkonen and Kärkkäinen [2008b], "Continuous reformulations and heuristics for the Euclidean travelling salesperson problem", ESAIM: Control, Opt. Calc. Var., doi:10.1051/cocv:2008056. The original publication is available at `www.esaim-cocv.org`.

we restrict ourselves to locally solving these penalised reformulations, by applying the perturbed Weiszfeld method applicable to finding "semi-critical" points of a sum of Euclidean distances from fixed points, perturbed by a concave function. Although applicable to the multisource Weber problem (providing a sort of dual of the K-means -style algorithm), it is unfortunately not applicable to the problem perturbed with the path length penalty. The algorithm is, however, applicable to the clustering formulation presented in Section 6.2, perturbed with the path length penalty. It is this latter reformulation we will use in our numerical experiments.

An (approximate) solution of such a continuous reformulation of the Euclidean TSP is not in practise – and not in theory either for big penalty parameters – a permutation of the original vertices. Therefore, along the course of studying these reformulations, we derive a heuristic that we use to "associate" the points of a solution with the original vertices. We also develop some other heuristics to reduce problem sizes, based on this heuristic and the clustering principle.

As for the applicability of our algorithms, we do not have any theoretical proofs of efficiency aside from partial convergence to "semi-critical points" (often local minima), and each step of the basic algorithm being $O(n^2)$ (consisting of n parallel Weiszfeld steps). On the experimental side, our method does seem to provide rather good results in quite few iterations for small problems. For bigger problems the performance, however, degrades considerably – there are, after all, many more local solutions then. A bigger penalty parameter value might help, but the algorithm we apply has a limit on its magnitude. Clustering heuristics that we develop, however, somewhat remedy the situation. Nevertheless, our numerical results are not remarkable compared to what is achievable with other (non-Euclidean) algorithms, as presented in Johnson and McGeoch [2002].

The primary contributions of this work are thus the reformulations that appear new and perhaps, with other methods applied to them, could provide better numerical results. The basic method based on the Weiszfeld algorithm is also new. Our clustering heuristics are related to the classic Karp clustering heuristic, Bentley's [1992] Fast Recursive Partitioning scheme, and Litke's [1984] clustering heuristic. The first two of these use a "hard-coded" partitioning approach until the clusters are small enough, after which the sub-problems in the cluster are solved either approximately or exactly. Our approach, by contrast, uses a more dynamic cluster configuration, as defined by a clustering problem objective function. Litke's method also uses an ad hoc dynamic clustering method. None of these methods incorporate TSP path length optimisation in the cluster calculation phase. Finally, our geometric penalisation approach bears some resemblance to various geometric neural net methods for the problem – see Johnson and McGeoch [1997] and the references therein – as well as the Lazy TSP of Polak and Wolansky [2007]. In this latter paper a formulation very similar to the first one of ours, but with squared distances, is analysed along with its convexification. This problem is also considered in Buttazzo and Stepanov [2004], in a wider measure-theoretic transport optimisation framework. The papers of Jones [1990] and Lerman [2003], considering the multiscale construction of paths cov-

ering infinite point sets or measures, also bear relationship to the geometric and clustering approach to the TSP.

The rest of this chapter is organised as follows: In Sections 7.2 and 7.3 we present our continuous reformulations. Then, in Section 7.4, we consider the sensitivity of the solutions of the penalised reformulations with respect to the solutions of the original problem, as the penalty parameter is varied. Section 7.5 considers heuristic approaches that could be used to improve or speed up results. Finally, in Section 7.6 we present and discuss the results of our numerical experiments.

7.2 First reformulation

Consider the Euclidean travelling salesperson problem

$$\min_{\sigma} \sum_{i=1}^{n} \|a_{\sigma i} - a_{\sigma(i+1)}\|, \tag{7.1}$$

where $\bar{a} \triangleq (a_1, \ldots, a_n) \in \mathbb{R}^{mn}$ are distinct vertices, also called cities, and σ is a permutation of the numbers $\{1, \ldots, n\}$, with $\sigma(n+1) \triangleq \sigma 1$. We shall henceforth use this identification without explicit mention. We denote by $\hat{\sigma}$ any of the optimal permutations that minimise (7.1). There are always at least n of these, every "shift" of a solution being one.

Let us now reformulate the problem as finding $\bar{y} \triangleq (y_1, \ldots, y_n)$ that solves

$$\min f_{\mathrm{TSP}}(\bar{y}) \triangleq \sum_{i=1}^{n} \|y_i - y_{i+1}\| \text{ subject to } y_i = a_{\sigma i} \text{ for some permutation } \sigma.$$

Here again we identify $y_{n+1} \triangleq y_1$. The qualification condition may be written as

$$f_{\mathrm{KM}}(\bar{y}; \bar{a}) \triangleq \sum_{i=1}^{n} \min_{j=1,\ldots,n} \|a_i - y_j\| = 0.$$

The function f_{KM} is precisely the multisource Weber problem (or "n-spatial medians") objective function, when the number of data points and cluster prototypes are equal; see Chapter 6. This function is diff-convex, as may be seen by rewriting $f_{\mathrm{KM}}(\bar{y}; \bar{a}) = f(\bar{y}; \bar{a}) - \nu_{\mathrm{KM}}(\bar{y}; \bar{a})$ with

$$f(\bar{y}; \bar{a}) \triangleq \sum_{i=1}^{n} \sum_{j=1}^{n} \|a_i - y_j\| \quad \text{and} \quad \nu_{\mathrm{KM}}(\bar{y}; \bar{a}) \triangleq \sum_{i=1}^{n} \max_{j=1,\ldots,n} \Big(\sum_{k \neq j} \|a_i - y_k\| \Big). \tag{7.2}$$

These considerations suggest relaxing problem (7.1) to the problem

$$\min_{\bar{y}} f_{\mathrm{KM}}(\bar{y}; \bar{a}) + \lambda f_{\mathrm{TSP}}(\bar{y}), \quad \lambda > 0, \tag{7.3}$$

or

$$\min_{\bar{y}} \left(\sum_{i=1}^{n} \min_{j=1,\ldots,n} \|a_i - y_j\| + \lambda \sum_{i=1}^{n} \|y_i - y_{i+1}\| \right).$$

Notice that for permutations σ of the vertices, $\bar{y} = \bar{a}_\sigma \triangleq (a_{\sigma 1}, \ldots, a_{\sigma n})$ are precisely all the global minimisers of (7.3) for $\lambda = 0$. The function f_{TSP} therefore acts as a perturbation to the multisource Weber problem, penalising such permutations that result in long paths. For small enough perturbation parameter λ, a minimiser $\hat{y}$ of (7.3) actually equals $\bar{a}_{\hat{\sigma}}$ for one of the optimal permutations $\hat{\sigma}$, as Theorem 7.1 below shows. For the proof we need first some preliminary results and definitions.

Definition 7.1. The vertices a_k ($k = 1, \ldots, n$) are *collinear* (on the line L) if there are vectors $z, v \in \mathbb{R}^m$ such that for the line $L \triangleq \mathbb{R}z + v$, $\{a_1, \ldots, a_n\} \subset L$. Otherwise the points are *non-collinear*.

Definition 7.2. Given a path/permutation σ, there is said to be a *degenerate angle* at the point $a_{\sigma k}$, if $(a_{\sigma(k+1)} - a_{\sigma k})^T (a_{\sigma(k-1)} - a_{\sigma k}) = \|a_{\sigma(k+1)} - a_{\sigma k}\| \|a_{\sigma(k-1)} - a_{\sigma k}\|$.

Since the collinear case is trivial, we will only consider the case of

Assumption 7.1. The vertices $a_k \in \mathbb{R}^m$ ($k = 1, \ldots, n$) are non-collinear and distinct.

The following result is well-known, but we provide the proof for completeness:

Lemma 7.1. *Suppose that Assumption 7.1 holds on the points $a_k \in \mathbb{R}^m$ ($k = 1, \ldots, n$). Then the points of an optimal path $\bar{a}_{\hat{\sigma}}$ form a simple closed curve. In particular, there are no degenerate angles.*

Proof. Assume without loss of generality that $\hat{\sigma}$ is the identity permutation. Suppose two (open) straight line segments of the path (a_k, a_{k+1}) and (a_i, a_{i+1}) with $i \neq k$, cross at a point c. Then replacing the former segments with (a_k, a_i) and (a_{k+1}, a_{i+1}), and reversing part of the remaining path, produces a valid path with one less crossing. Now

$$\begin{aligned} \|a_k - a_i\| + \|a_{k+1} - a_{i+1}\| &\leq \|a_k - c\| + \|a_i - c\| + \|a_{k+1} - c\| + \|a_{i+1} - c\| \\ &= \|a_k - a_{k+1}\| + \|a_i - a_{i+1}\|, \end{aligned}$$

with the inequality strict if c does not lie on one (and then both) of the segments (a_k, a_i) or (a_{k+1}, a_{i+1}). Thus the path can in that case be improved by removing the crossing.

If $c \in (a_k, a_i) \cap (a_{k+1}, a_{i+1})$, then these points are collinear, and (a_k, a_{k+1}) or (a_i, a_{i+1}) contains an endpoint of the other; say $a_k \in [a_i, a_{i+1}]$, the other cases being analogous. The path can therefore visit a_k during this segment, not increasing the cost. Furthermore, if this segment is part of the optimal path, the smaller problem with a_k removed will have equal optimal path length. If removing a_k does not improve the path length by going from a_{k-1} directly to a_{k+1}, it must be that a_{k-1}, a_{k+1}, a_i and a_{i+1} are collinear. Therefore, if recursively applying the argument never improves the path, all the points must be collinear. This is in contradiction to our assumptions. □

Note that $\partial\|\cdot - a\|(a) = \mathbb{B}(0,1)$, yielding (by local convexity) that $\partial f_{\mathrm{KM}}(\bar{a}_\sigma; \bar{a}) = \prod_{i=1}^n \mathbb{B}(0,1)$ when the points are distinct.

Theorem 7.1.

(i) For $\lambda \in (0, 1/2]$, every global minimiser $\bar{y}$ of (7.3), is a permutation of $\bar{a}$.

(ii) For $\lambda \in (0, 1/2)$, global minimisers of (7.3), coincide with optimal TSP paths $\bar{a}_{\hat{\sigma}}$; the same holds for $\lambda = 1/2$ under Assumption 7.1.

(iii) However, for every permutation σ, $\bar{a}_\sigma$ is a strict local minimiser of (7.3) for $\lambda \in [0, 1/2)$ and a (possibly non-strict) local minimiser for $\lambda = 1/2$.

Proof. Let $\bar{y} = (y_1, \ldots, y_n) \in \mathbb{R}^{mn}$ be arbitrary. Suppose that for some y_j ($j = 1, \ldots, n$) the following property holds: for every a_k ($k = 1, \ldots, n$) and some $i(k) \neq j$, $\|a_k - y_{i(k)}\| \leq \|a_k - y_j\|$. The point y_j then does not contribute to f_{KM}, and we may assume that it lies on the straight line segment from y_{j-1} to y_{j+1}, for otherwise the cost could be decreased by making this alteration. We may in fact freely move y_j on the path composed of the remaining points y_i ($i \neq j$). Therefore we can arrange the points in such a way that whenever y_j minimises $i \mapsto \|a_k - p_i\|$ for N_j points a_k, then the multiplicity of y_i with $y_i = y_j$ is also N_j.

The (possibly collinear) case with $\lambda \in (0, 1/2)$. Let then y_j minimise $i \mapsto \|a_k - y_i\| > 0$. We may then alter $\bar{y}$ by assigning $y_j \mapsto a_k$, actually decreasing the cost. This follows from the following two observations. Firstly, a) by the previous alterations, if y_j is a minimiser of the distance for another $a_\ell \neq a_k$, then there is also another $y_i = y_j$ for which this holds. Therefore $\min_i \|a_\ell - y_i\|$ is not increased. Secondly, b) for $\lambda \in (0, 1/2)$, we have

$$\lambda\|y_{j-1} - a_k\| < \lambda\|y_{j-1} - y_j\| + \frac{1}{2}\|y_j - a_k\| \tag{7.4}$$

and similarly for y_{j+1}. Thus the increase in the length of the path $(y_1, \ldots, y_n, y_1)$ is consumed by the decrease of $\min_j \|a_k - y_j\|$ to zero.

We have therefore showed that for $\lambda \in (0, 1/2)$, only the points $\bar{a}_\sigma$ for permutations σ can be global minimisers. Obviously the actual global minimisers correspond to the permutations that minimise f_{TSP}.

However, $0 \in \operatorname{int} \partial(f_{\mathrm{KM}}(\cdot; \bar{a}) + \lambda f_{\mathrm{TSP}})(\bar{a}_\sigma)$ because $\partial f_{\mathrm{KM}}(\bar{a}_\sigma; \bar{a}) = \prod_{i=1}^n \mathbb{B}(0,1)$ as already noted, and

$$\begin{aligned}\nabla_{y_i} f_{\mathrm{TSP}}(\bar{a}_\sigma) &= \nabla_{y_i}(\|y_i - a_{\sigma(i-1)}\| + \|y_i - a_{\sigma(i+1)}\|)(a_{\sigma i}) \\ &= \frac{a_{\sigma i} - a_{\sigma(i+1)}}{\|a_{\sigma i} - a_{\sigma(i+1)}\|} + \frac{a_{\sigma i} - a_{\sigma(i-1)}}{\|a_{\sigma i} - a_{\sigma(i-1)}\|}\end{aligned}$$

for $\lambda \in (0, 1/2)$. By the local convexity of f_{KM} in a neighbourhood of $\bar{a}_\sigma$, strict local optimality follows.

When $\lambda = 1/2$, we still have $0 \in \partial(f_{\mathrm{KM}}(\cdot; \bar{a}) + \lambda f_{\mathrm{TSP}})(\bar{a}_\sigma)$. Thus local optimality follows from local convexity. For an optimal permutation $\hat{\sigma}$, by Lemma 7.1 we must in fact have $\|\nabla_i f_{\mathrm{TSP}}(\bar{a}_{\hat{\sigma}})\| < 2$, wherefore strict local optimality still holds. It remains to prove global optimality for this case.

The non-collinear case with $\lambda = 1/2$. Let again y_j minimise $i \mapsto \|a_k - y_i\| > 0$. The inequality in (7.4) still holds as non-strict. In fact, when it holds as equality for both $j-1$ and $j+1$, all the points y_j, y_{j-1}, y_{j+1} and a_k must lie on a line L, such that in one of the natural orders $\prec$ of L, $a_k \prec y_j$, $y_j \prec y_{j+1}$, and $y_j \prec y_{j-1}$. As before, we may then move y_j to $y'_j \triangleq a_k$, not increasing the cost. Since $\|a_k - y_j\| > 0$ was minimal, y'_j can equal neither y_{j-1} nor y_{j+1}. Therefore there is a degenerate angle in the altered path at y'_j. Now, if some y_i is not on L, Lemma 7.1 applied to the points $y_1, \ldots, y'_j, \ldots, y_n$ (duplicates removed) shows that the path cannot be optimal.

The possibility then remains that all the points y_i are on L. By the non-collinearity assumption, there is some a_k that is not on L. But now (7.4) holds strictly for the y_j minimising $i \mapsto \|a_k - y_i\| > 0$. Therefore the cost can be decreased as before. □

Corollary 7.1. *Finding a point arbitrarily close to a minimiser of problem* (7.3) *is NP-hard for $\lambda \in (0, 1/2]$ when the vertices are non-collinear (with rational coordinates). Consequently, we have another proof that finding minimisers of diff-convex functions is NP-hard.*

Proof. We can always assume that $\|a_k - a_\ell\| \geq 1$ ($k \neq \ell$), because scaling does not alter $\hat{\sigma}$. Suppose then that for problem (7.3) and a given $\epsilon > 0$, we were able to find in time polynomial in n (but not in ϵ), a point $\hat{y}$ with $\|\hat{y}_i - a_{\hat{\sigma}i}\| < \epsilon$ ($i = 1, \ldots, n$) for some $\hat{\sigma}$. Then, taking $\epsilon = 1/3$, we could uniquely assign each $\hat{y}_i$ to $a_{\hat{\sigma}i}$ in polynomial time. But this means we could solve the original NP-hard Euclidean TSP problem (7.1) in polynomial time. □

For small enough λ, a good enough approximate solution should therefore identify the solution of problem (7.1), there being a unique distance-minimising assignment of each y_j to a_k. For parameters greater than the threshold value of λ, one could look for a permutation σ for which $\bar{a}_\sigma$ closely matches $\bar{y}$, for example by following the method used in the proof of Theorem 7.1. Deciding how to optimally assign equal points y_j to the corresponding vertices in that method, can of course be expensive in itself.

The benefit from using a bigger λ comes from the local minima starting to disappear as the objective function becomes "more convex", and therefore possibly easier to minimise. For very big λ, the global minimisers also drift far from the sought solution, however: the study of this sensitivity is the topic of Section 7.4.

By the diff-convexity, one could thus try to solve problem (7.1) by (approximately) solving a penalised version (7.3) by methods of global optimisation, such as outer approximation methods [see, e.g., Horst and Pardolos, 1995]. As stated, we are, however, interested in applying the somewhat more lightweight perturbed Weiszfeld method from Chapter 5 to the problem. Unfortunately, the present model does not exactly fit within the class of problems considered in Chapter 5, for which we have partial convergence proofs. The problem is that $f_{\mathrm{TSP}}(\bar{y}) - \nu_{\mathrm{KM}}(\bar{y}; \bar{a})$ is not concave.

7.3 Second reformulation

We are thus led to seek for another way to formulate the condition $y_i = a_{\sigma i}$, that would fit within the above-mentioned class of problems. Given the observed relationship to the K-means clustering problem, a natural candidate is based on the multi-objective clustering problem formulated in Chapter 6. The problem then becomes

$$\min_{\bar{y}} f_{\mathrm{MO}}(\bar{y};\bar{a}) + \lambda f_{\mathrm{TSP}}(\bar{y}), \quad \lambda > 0, \tag{7.5}$$

where

$$f_{\mathrm{MO}}(\bar{y};\bar{a}) \triangleq f(\bar{y};\bar{a}) - \nu_{\mathrm{MO}}(\bar{y})$$

is in structure similar to f_{KM}: the function ν_{KM} has merely been replaced with

$$\nu(\bar{y}) \triangleq \nu_{\mathrm{MO}}(\bar{y}) \triangleq \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|y_i - y_j\|.$$

We have fixed the factor $1/2$ already at this point for simplicity; in Chapter 6 this may vary up to $(1/2)n/(K-1)$, with $K = n$ in the present case, while ensuring level-boundedness of the objective function.

This time, the function $f_{\mathrm{TSP}}(\bar{y}) - \nu(\bar{y})$ is concave for $\lambda \in [0,1]$, because $\nu(\bar{y})$ contains all the terms $\|y_i - y_{i+1}\|$ in the sum expression.

The permutations $\bar{a}_\sigma$ are strict local minimisers of $f_{\mathrm{MO}}(\cdot;\bar{a})$, as

$$\nabla\nu(\bar{a}_\sigma) \in \operatorname{int}\partial f(\bar{a}_\sigma;\bar{a}) = \nabla\nu(\bar{a}_\sigma) + \prod_{i=1}^{n}\operatorname{int}\mathbb{B}(0,1). \tag{7.6}$$

This follows from

$$\nabla_{y_i}\nu(\bar{a}_\sigma) = \sum_{j\neq i}\nabla_{y_i}\|y_i - a_{\sigma j}\|(a_{\sigma i}), \quad \partial_{y_i}f(\bar{a}_\sigma) = \sum_{k}\partial_{y_i}\|y_i - a_k\|(a_{\sigma i}),$$

with the difference $\mathbb{B}(0,1)$ coming from $\sigma i = k$.

As before, we have the inclusion $\partial f_{\mathrm{TSP}}(\bar{a}_\sigma) \subset \mathbb{B}(0,1)$, strict at $\sigma = \hat{\sigma}$ under Assumption 7.1 by Lemma 7.1. Therefore, all the points $\bar{a}_\sigma$ are strict local minimisers for $\lambda \in (0,1/2)$, and $\bar{a}_{\hat{\sigma}}$ for $\lambda = 1/2$ as well. As for global optimality, we have:

Theorem 7.2. *Suppose the points a_i $(i = 1,\ldots,n)$ are distinct. Then,*

(i) The global minimisers of $f_{\mathrm{MO}}(\cdot;\bar{a})$ are exactly $\bar{a}_\sigma$ for all σ.

(ii) There exists a $\hat{\lambda} > 0$, such that the minimisers of $f^{\lambda}_{\mathrm{MOTSP}} \triangleq f_{\mathrm{MO}}(\cdot;\bar{a}) + \lambda f_{\mathrm{TSP}}$ are exactly the optimal TSP paths $\bar{a}_{\hat{\sigma}}$ for $\lambda \in (0,\hat{\lambda})$.

We begin the proof with a few lemmas. For the case $m > 1$, we will use the following extension (to strict inequalities) of a reduction theorem of Levi [see Mitrinović, 1970, p. 175].

Lemma 7.2. *Let $k_i \in \mathbb{R}$ and $\rho_{ij} \in \mathbb{R}$, $i = 1, \ldots, K$, $j = 1, \ldots, N$. Suppose that for all $\bar{x} = (x_1, \ldots, x_N) \in \mathbb{R}^N$, we have*

$$\sum_{i=1}^{K} k_i |\rho_{i1} x_1 + \cdots + \rho_{iN} x_N| \geq 0, \tag{7.7}$$

and let C be the cone of $\bar{x}$s, on which this inequality is strict. Then for all $\bar{y} = (y_1, \ldots, y_N) \in \mathbb{R}^{mN}$,

$$\sum_{i=1}^{K} k_i \|\rho_{i1} y_1 + \cdots + \rho_{iN} y_N\| \geq 0.$$

Furthermore, this inequality is strict on the cone C' where

$$A(\bar{y}) \triangleq \{b \in \mathbb{R}^m \mid \|b\| = 1, (b^T y_1, \ldots, b^T y_N) \in C\}$$

has positive Lebesgue measure on the unit sphere.

Proof. Let $\xi_i \triangleq \rho_{i1} y_j + \cdots + \rho_{iN} y_N$. Then for some constant $C_m > 0$,

$$\begin{aligned}\sum_{i=1}^{K} k_i \|\xi_i\| / C_m = \sum_{i=1}^{K} k_i \|\xi_i\| \int_{\|b\|=1} |b^T \xi_i / \|\xi_i\|| db &= \int_{\|b\|=1} \sum_{i=1}^{K} k_i |b^T \xi_i| db \\ &= \int_{\|b\|=1} \sum_{i=1}^{K} k_i |\rho_{i1} x_j(b) + \cdots + \rho_{iN} x_N(b)| db \geq 0,\end{aligned}$$

where $x_j(b) \triangleq b^T y_j$. As the area integrated over includes $A(\bar{y})$, the claim on strictness of the inequality follows. □

Lemma 7.3. *Suppose the points a_i $(i = 1, \ldots, n)$ are distinct, and define $r_\sigma(\bar{y}) \triangleq 2 \sum_i \|y_i - a_{\sigma i}\|$. Then for all $\lambda \in [0, 1/2)$, and permutations σ, there exist neighbourhoods E_σ^λ of $\bar{a}_\sigma$, where*

$$f_{\mathrm{MO}}(\bar{y}; \bar{a}) - f_{\mathrm{MO}}(\bar{a}_\sigma; \bar{a}) \geq \lambda r_\sigma(\bar{y}), \quad \bar{y} \in E_\sigma^\lambda. \tag{7.8}$$

When $m = 1$, $\bigcup_\sigma E_\sigma^\lambda = \mathbb{R}^{nm}$ (i.e., the whole space), and when $\lambda = 0$, $E_\sigma^0 = \mathbb{R}^{nm}$. In both of these cases, the inequality holds strictly when $\bar{y} \neq \bar{a}_\sigma$ for all σ.

Proof. We have

$$\begin{aligned}f(\bar{y}; \bar{a}) - f(\bar{a}_\sigma; \bar{a}) &\geq \max \partial f(\bar{a}_\sigma; \bar{a})^T (\bar{y} - \bar{a}_\sigma) \\ &= \max[\prod_{i=1}^{n} \mathbb{B}(0,1) + \nabla \nu(\bar{a}_\sigma)]^T (\bar{y} - \bar{a}_\sigma) \\ &= (1/2) r_\sigma(\bar{y}) + \nabla \nu(\bar{a}_\sigma)^T (\bar{y} - \bar{a}_\sigma),\end{aligned}$$

where the subdifferential is calculated as for (7.6), and the last equality follows from the expression $\|x\| = \max\{z^T x \mid z \in \mathbb{B}(0,1)\}$ for $x \in \mathbb{R}^m$. Likewise,

$$\nu(\bar{a}_\sigma) - \nu(\bar{y}) \geq \partial \nu(\bar{y})^T (\bar{a}_\sigma - \bar{y}).$$

Therefore

$$f_{\mathrm{MO}}(\bar{y}) - f_{\mathrm{MO}}(\bar{a}_\sigma) = [f(\bar{y};\bar{a}) - f(\bar{a}_\sigma;\bar{a})] + [\nu(\bar{a}_\sigma) - \nu(\bar{y})]$$
$$\geq (1/2) r_\sigma(\bar{y}) - \min[\partial\nu(\bar{y}) - \nabla\nu(\bar{a}_\sigma)]^T(\bar{y} - \bar{a}_\sigma).$$

By monotonicity $[\partial\nu(\bar{y}) - \nabla\nu(\bar{a}_\sigma)]^T(\bar{y} - \bar{a}_\sigma) \geq 0$. The problem is now to bound

$$L \triangleq \min[\partial\nu(\bar{y}) - \nabla\nu(\bar{a}_\sigma)]^T(\bar{y} - \bar{a}_\sigma) \leq (1/2 - \lambda) r_\sigma(\bar{y}). \tag{7.9}$$

But, since the a_i are distinct, ν is continuously differentiable[1] in some neighbourhood of each a_σ. Now, we approximate

$$L \leq \sum_{i=1}^{n} \|[\nabla\nu(\bar{y}) - \nabla\nu(\bar{a}_\sigma)]_i\| \|y_i - a_{\sigma i}\| \leq \max_j \|\nabla_j\nu(\bar{y}) - \nabla_j\nu(\bar{a}_\sigma)\| r_\sigma(\bar{y})/2.$$

From this we see that some neighbourhoods E_σ^λ of $\bar{a}_\sigma$ can be found, where the maximum term is small enough for (7.9) to hold.

Now, if $m = 1$, there actually exists for each $\bar{y}$ a permutation σ, for which the left hand side of (7.9) is zero. Therefore, for all $\lambda \leq 1/2$, (7.9) and then (7.8) hold, and $\bigcup_\sigma E_\sigma^\lambda = \mathbb{R}$. To see this, recall that where ν is differentiable (i.e., $y_i \neq y_j$ for $i \neq j$),

$$\nabla_{y_i}\nu(\bar{y}) = \sum_{j \neq i}^{n} \frac{y_i - y_j}{\|y_i - y_j\|}.$$

In the $m = 1$ case the terms summed over are ± 1, indicating the direction y_j faces from y_i on the real line. But the set of these numbers over all i then uniquely determines the order of the y_i on the real line, and consequently a permutation σ, for which $\nabla\nu(\bar{y}) = \nabla\nu(\bar{a}_\sigma)$. In the non-differentiable case, $y_i = y_j$ for some $i \neq j$. In this case we can arbitrarily decide on the order, and choose the corresponding signs ± 1 from $\partial_{(y,y')}\|y - y'\|(y,y) = \{(z,-z) \mid z \in \mathbb{B}(0,1)\}$.

The claim on strictness of the inequality (7.8) in the $m = 1$ case follows from the non-strict variant, since $E_\sigma^{1/2}$ cover the whole space, and $(1/2) r_\sigma(\bar{y}) > \lambda r_\sigma(\bar{y})$ when $\bar{y} \neq a_\sigma$ and $\lambda < 1/2$.

Now, if $\lambda = 0$ (and still $m = 1$), the right hand side of (7.8) is zero, and independent of σ. We have also previously shown that for every $\bar{y}$, the inequality holds for some σ. But since $2\nu(\bar{a}_\sigma) = f(\bar{a}_\sigma;\bar{a})$ and $f_{\mathrm{MO}}(\bar{a}_\sigma) = f(\bar{a}_\sigma;\bar{a}) - \nu(\bar{a}_\sigma) = \nu(\bar{a}_\sigma) = \nu(\bar{a})$ does not depend on σ, actually

$$f_{\mathrm{MO}}(\bar{y};\bar{a}) - f_{\mathrm{MO}}(\bar{a}_\sigma;\bar{a}) = f(\bar{y};\bar{a}) - \nu(\bar{y}) - \nu(\bar{a}) \geq 0 \quad \text{for all } \sigma \text{ and } \bar{y}. \tag{7.10}$$

Therefore, in the $m = 1$ case, $E_\sigma^0 = \bigcup_{\sigma'} E_{\sigma'}^0 = \mathbb{R}$ for all σ.

Finally, suppose $m > 1$ and $\lambda = 0$. Since (7.10) is of the form (7.7) with $\bar{x} = (\bar{y},\bar{a})$ when $y_i, a_k \in \mathbb{R}$, we may apply Lemma 7.2 with $\bar{y} = (\bar{y},\bar{a})$ when $y_i, a_k \in \mathbb{R}^m$ to obtain that (7.10) holds generally. For the strict inequality, to show that $A(\bar{y},\bar{a})$ has positive measure, choose the projection b in Lemma 7.2 so that (7.8) holds strictly, i.e., at least for some i, $b^T y_i \neq b^T a_k$ for all k. This can be done if

[1] Twice actually, so we could alternatively apply the mean value theorem.

$y_i \neq a_k$ for all k, because the set of projections with $b^T y_i = b^T a_k$ is then finite. By continuity, the same holds in a neighbourhood of positive measure of the chosen points and projection. Therefore $A(\bar{y}, \bar{a})$ has positive measure. □

Proof of Theorem 7.2. Lemma 7.3 with $\lambda = 0$ proves claim (i).

As for claim (ii), since f_{MO} is continuous and level-bounded (as noted above), the cluster points of minimisers $\hat{y}_{[\lambda]}$ of $f^{\lambda}_{\text{MOTSP}}$ with $\lambda \searrow 0$, must be those of $f_{\text{MO}} = f^{0}_{\text{MOTSP}}$ [Rockafellar and Wets, 1998, Theorem 1.17]. Since there are finitely many permutations σ, there is a constant c dependent on $\bar{a}$, such that $f_{\text{TSP}}(\bar{a}_\sigma) \geq c + f_{\text{TSP}}(\bar{a}_{\hat{\sigma}})$ for non-optimal σ. Therefore the cluster points must be the optimal TSP paths $\bar{a}_{\hat{\sigma}}$.

To show the existence of the threshold on λ, choose an arbitrary $\tilde{\lambda} \in (0, 1/2)$. There must now exist $\hat{\lambda} \leq \tilde{\lambda}$, such that $\hat{y}_{[\lambda]} \in E \triangleq \bigcup_{\hat{\sigma}} E^{\tilde{\lambda}}_{\hat{\sigma}}$ for $\lambda \in (0, \hat{\lambda})$. If this were not so, we could find a cluster point outside E, in contradiction to previously established results.

Now, apply

$$\begin{aligned} \|a_{\sigma i} - a_{\sigma(i+1)}\| &= \|a_{\sigma i} - y_i + y_i - y_{i+1} + y_{i+1} - a_{\sigma(i+1)}\| \\ &\leq \|a_{\sigma i} - y_i\| + \|y_i - y_{i+1}\| + \|y_{i+1} - a_{\sigma(i+1)}\|, \end{aligned} \tag{7.11}$$

to yield

$$f_{\text{TSP}}(\bar{a}_\sigma) - f_{\text{TSP}}(\bar{y}) = \sum_i (\|a_{\sigma i} - a_{\sigma(i+1)}\| - \|y_i - y_{i+1}\|) \leq 2 \sum_i \|y_i - a_{\sigma i}\|. \tag{7.12}$$

Combined with (7.8), we therefore have

$$\begin{aligned} f^{\lambda}_{\text{MOTSP}}(\bar{y}) - f^{\lambda}_{\text{MOTSP}}(\bar{a}_{\hat{\sigma}}) &= f_{\text{MO}}(\bar{y}) - f_{\text{MO}}(\bar{a}_{\hat{\sigma}}) + \lambda(f_{\text{TSP}}(\bar{y}) - f_{\text{TSP}}(\bar{a}_{\hat{\sigma}})) \\ &\geq (\tilde{\lambda} - \lambda) r_{\hat{\sigma}}(\bar{y}), \end{aligned}$$

whenever $\lambda \in [0, \tilde{\lambda})$ and $\bar{y} \in E^{\tilde{\lambda}}_{\hat{\sigma}}$. This says that for $\lambda \in (0, \hat{\lambda})$, we must have $\hat{y}_{[\lambda]} = \bar{a}_{\hat{\sigma}}$ for some $\hat{\sigma}$. □

Corollary 7.2. *Either (or both) the calculation of $\hat{\lambda}$ is NP-hard, or the problem* (7.5) *is NP-hard for $\lambda \in (0, \hat{\lambda})$ (and non-collinear vertices $\bar{a}$).*

Proof. Identical to Corollary 7.1. □

Remark 7.1. Actually, the upper bound $\hat{\lambda}$ is not strict under Assumption 7.1, for (7.11) is strict for some $i \in \{1, \ldots, n\}$. Suppose it were not. Then all the vectors $a_{\sigma i} - y_i$, $y_i - y_{i+1}$, and $y_{i+1} - a_{\sigma(i+1)}$ would point in the same direction, for all i. But this cannot be unless both y_i and y_{i+1} are collinear with $a_{\sigma i}$ and $a_{\sigma(i+1)}$. Therefore all the four points are collinear. But likewise y_{i+1} and $a_{\sigma(i+1)}$ are collinear also with y_{i+2} and $a_{\sigma(i+2)}$. By extension, all the points $y_1, \ldots, y_n$ and, in particular, $a_1, \ldots, a_n$ are collinear, which violates our assumptions.

Lemma 7.3 also contains the following interesting special cases, obtained with $\lambda = 0$, stated here separately:

Corollary 7.3. *For any points $a_1, \ldots, a_n \in \mathbb{R}^m$, and $y_1, \ldots, y_n \in \mathbb{R}^m$, it holds that*

$$\sum_{i=1}^{n}\sum_{k=1}^{n}\|y_i - a_k\| \geq \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|y_i - y_j\| + \frac{1}{2}\sum_{k=1}^{n}\sum_{\ell=1}^{n}\|a_k - a_\ell\|.$$

In particular, when $y_i = -a_i$,

$$\sum_{k=1}^{n}\sum_{\ell=1}^{n}\|a_k + a_\ell\| \geq \sum_{k=1}^{n}\sum_{\ell=1}^{n}\|a_k - a_\ell\|.$$

Proof. This follows from the equivalence and inequality in (7.10). □

7.4 Sensitivity analysis

Here we provide some sensitivity results for our penalised reformulations of the Euclidean TSP. This is in order to understand how the solutions vary, as the penalty parameter λ varies above $1/2$ or $\hat{\lambda}$, and to justify the use of values higher than this threshold. We define $f^{\lambda}_{\mathrm{KMTSP}} \triangleq f_{\mathrm{KM}}(\cdot;\bar{a}) + \lambda f_{\mathrm{TSP}}$. Note that this function is locally convex at $\bar{a}_{\hat{\sigma}}$, so that the convex subdifferential is defined there. For the reformulation based on the multisource Weber problem, we then get the following theorem.

Theorem 7.3. *Suppose Assumption 7.1 holds. Let $\gamma \geq 0$, $\lambda_0 \in (0, 1/2]$, and $\lambda \geq \lambda_0$. Suppose $D \cap \arg\min f^{\lambda_0}_{\mathrm{KMTSP}} \neq \emptyset$, and that $\hat{y} \in \gamma\text{-}\arg\min_D f^{\lambda}_{\mathrm{KMTSP}}$. Denote $\eta \triangleq f_{\mathrm{TSP}}(\bar{a}_{\hat{\sigma}}) - \min_D f_{\mathrm{TSP}}$ (this value does not depend on the choice of $\hat{\sigma}$), and $C_{\hat{\sigma}} \triangleq \partial f^{\lambda_0}_{\mathrm{KMTSP}}(\bar{a}_{\hat{\sigma}})$. Then,*

(i) If for some $\hat{\sigma}$, also $\hat{y} \in \bar{a}_{\hat{\sigma}} + \prod_{i=1}^{n}\mathbb{B}(0, \delta_{\hat{\sigma}i})$ for $\delta_i \triangleq \min_{j\neq i}\|a_i - a_j\|/2$, we actually have

$$\hat{y} \in \bar{a}_{\hat{\sigma}} + ((\lambda - \lambda_0)\eta + \gamma)C^{\circ}_{\hat{\sigma}}, \tag{7.13}$$

with the set on the right bounded.

(ii) Suppose (7.13) holds and $\lambda \leq \lambda_0 + \min_i(\delta_{\hat{\sigma}i} - \gamma M_i)(\eta M_i)^{-1}$ with $M_i = \max_{\bar{x}\in C^{\circ}_{\hat{\sigma}}}\|x_i\|$. Then $\hat{y} \in \bar{a}_{\hat{\sigma}} + \prod_{i=1}^{n}\mathbb{B}(0, \delta_{\hat{\sigma}i})$.

(iii) There exists a finite index set $\mathcal{T}$, closed sets E_t, compact sets C_t, points $\bar{q}_t$, and constants $c_t \in [0, f^{\lambda_0}_{\mathrm{KMTSP}}(\bar{q}_t) - f^{\lambda_0}_{\mathrm{KMTSP}}(\bar{a}_{\hat{\sigma}})]$, such that C°_t is bounded, and for some $t \in \mathcal{T}$,

$$\hat{y} \in E_t \cap \big(\bar{q}_t + ((\lambda - \lambda_0)\eta + \gamma - c_t)C^{\circ}_t\big).$$

(iv) For $\hat{y} \in E_t$ in (iii), we must have $(\lambda - \lambda_0)\eta + \gamma \geq f^{\lambda_0}_{\mathrm{KMTSP}}(\bar{q}_t) - \min f^{\lambda_0}_{\mathrm{KMTSP}}$.

Proof. (i) We have

$$\sup_D\big(f^{\lambda_0}_{\mathrm{KMTSP}} - f^{\lambda}_{\mathrm{KMTSP}}\big) - \inf_{\bar{a}_{\hat{\sigma}}}\big(f^{\lambda_0}_{\mathrm{KMTSP}} - f^{\lambda}_{\mathrm{KMTSP}}\big)$$
$$= \sup_D(\lambda_0 - \lambda)(f_{\mathrm{TSP}} - f_{\mathrm{TSP}}(\bar{a}_{\hat{\sigma}})) = \eta' \triangleq (\lambda - \lambda_0)\eta,$$

wherefore η' provides the distance between the functions on D, needed for the application of Corollary 2.5. We want to exploit local convexity in doing so.

When $\hat{y} \in \bar{a}_{\hat{\sigma}} + \prod_{i=1}^{n} \mathbb{B}(0, \delta_i)$, the choice of δ_i forces the distance $\|y_i - a_{\hat{\sigma}i}\|$ to be minimal for both y_i and $a_{\hat{\sigma}i}$ (against alternatives of the other), so that both $\hat{y}$ and $\bar{a}_{\hat{\sigma}}$ belong to a neighbourhood $E_{\bar{a}_{\hat{\sigma}}}$ on which f_{KM} is locally convex. Therefore $f_0 \triangleq f^{\lambda_0}_{\mathrm{KMTSP}}$ is also convex in this neighbourhood, and we may actually take $\nu = 0$ and $f = f_0|E_{\bar{a}_{\hat{\sigma}}}$ to be the restriction of f_0 (defined to be $+\infty$ outside $E_{\bar{a}_{\hat{\sigma}}}$) in Corollary 2.5, akin to Example 2.6. This provides the desired sensitivity result with $\epsilon' = \bar{\epsilon} = 0$, as then $A \triangleq \{\bar{a}_{\hat{\sigma}}\} \subset E_{\bar{a}_{\hat{\sigma}}} \cap D \subset \mathbb{R}^{nm} = D_{\bar{a}_{\hat{\sigma}}}(\epsilon')$, and

$$\begin{aligned}\hat{y} \in U_{\bar{a}_{\hat{\sigma}}}(\eta' + \gamma, \epsilon') &= \bar{a}_{\hat{\sigma}} + \bigcup_{\epsilon_\nu \in [0, \epsilon']} \bigcap_{\epsilon \in [\epsilon_\nu, \bar{\epsilon})} (\eta' + \gamma + \epsilon - \epsilon_\nu)(\partial_\epsilon (f_0|E_{\bar{a}_{\hat{\sigma}}}))^\circ(\bar{a}_{\hat{\sigma}}) \\ &= \bar{a}_{\hat{\sigma}} + (\eta' + \gamma) C^\circ_{\hat{\sigma}}.\end{aligned}$$

The simplified expression of $U_{\bar{a}_{\hat{\sigma}}}$ above (cf. Lemma 2.4) is justified, because $0 \in \operatorname{int} C_{\hat{\sigma}}$, which also implies that $C^\circ_{\hat{\sigma}}$ is bounded. To see this, since $\partial f_{\mathrm{KM}}(\bar{a}_{\hat{\sigma}}) = \prod_{i=1}^{n} \mathbb{B}(0,1)$, it suffices to show that $\beta_i < 2$ for

$$\beta_i \triangleq \left\| \frac{a_{\hat{\sigma}i} - a_{\hat{\sigma}(i-1)}}{\|a_{\hat{\sigma}i} - a_{\hat{\sigma}(i-1)}\|} + \frac{a_{\hat{\sigma}i} - a_{\hat{\sigma}(i+1)}}{\|a_{\hat{\sigma}i} - a_{\hat{\sigma}(i+1)}\|} \right\| = \|\nabla_i f_{\mathrm{TSP}}(\bar{a}_{\hat{\sigma}})\|.$$

But $\beta_i \leq 2$, and equality can only happen when there is a degenerate angle between $a_{\hat{\sigma}(i-1)}$, $a_{\hat{\sigma}i}$, and $a_{\hat{\sigma}(i+1)}$. By Lemma 7.1 and Assumption 7.1 this cannot happen for optimal permutations $\hat{\sigma}$.

(ii) The condition $\lambda \leq \lambda_0 + \min_i (\delta_{\hat{\sigma}i} - \gamma M_i)(\eta M_i)^{-1}$ is equivalent to $(\lambda - \lambda_0)\eta + \gamma \leq \delta_{\hat{\sigma}i}/M_i$ for all $i = 1, \ldots, n$. Thus (7.13) says $\hat{y} \in \bar{a}_{\hat{\sigma}} + \{(\delta_{\hat{\sigma}1} x_1/M_1, \ldots, \delta_{\hat{\sigma}n} x_n/M_n) \mid \bar{x} \in C^\circ_{\hat{\sigma}}\} \subset \bar{a}_{\hat{\sigma}} + \prod_{i=1}^{n} \mathbb{B}(0, \delta_{\hat{\sigma}i})$ by the definition of M_i.

(iii) Notice that $f_0 = f^{\lambda_0}_{\mathrm{KMTSP}}$ is convex on a finite family $\{E_t \mid t \in \mathcal{T}\}$ of closed sets – corresponding to different associations of the y_i to a_k (possibly multiple/empty) – that fill the entire space. On these regions f_{KM} is equal to some convex function $f^t : \bar{y} \mapsto \sum_{k=1}^{n} \|a_k - y_{i(k,t)}\|$ for some association $i(k,t)$ (not necessarily a permutation). Let $\bar{q}_t$ be a minimiser of $f^t_{\mathrm{TSP}} \triangleq f^t + \lambda_0 f_{\mathrm{TSP}}$, not necessarily in E_t. The subdifferential of f^t_{TSP} may be a singleton at $\bar{q}_t$, and thus not provide much information. But we can use a more informative approximate subdifferential containing 0 in its interior and thus with bounded polar, as follows.

The function f^t_{TSP} is level-bounded: Suppose $\|\bar{z}\| = 1$. Then the triangle inequality gives $f^t_{\mathrm{TSP}}(\bar{q}_t + \alpha\bar{z})/\alpha \geq \sum_{k=1}^{n} \|z_{i(k,t)}\| + \lambda_0 f_{\mathrm{TSP}}(\bar{z}) - f^t_{\mathrm{TSP}}(\bar{q}_t)/\alpha$. If $f_{\mathrm{TSP}}(\bar{z}) < \delta$, we must have $\|z_k - z_i\| < \delta$ for all $k, i = 1, \ldots, n$. But then, for small enough $\delta > 0$, by $\|\bar{z}\| = 1$ each $\|z_i\|$ must be close to $1/\sqrt{n}$. Thus $\sum_{k=1}^{n} \|z_{i(k,t)}\| + f_{\mathrm{TSP}}(\bar{z})$ is bounded from below on $\{\|\bar{z}\| = 1\}$ by some value greater than zero. Therefore, for big enough $\alpha > 0$, $f^t_{\mathrm{TSP}}(\bar{q}_t + \alpha\bar{z})/\alpha$ is greater than some constant, and hence f^t_{TSP} is level-coercive and then level-bounded. Thus by Lemma A3.2 in Appendix 3, $0 \in \operatorname{int} \partial_{\epsilon_t} f^t_{\mathrm{TSP}}(\bar{q}_t)$ for $\epsilon_t > 0$.

We may assume that $\bar{q}_t \neq \bar{a}_{\hat{\sigma}}$, for otherwise claim i) provides the result. Let then $C_t \triangleq \partial_{\epsilon_t} f^t_{\mathrm{TSP}}(\bar{q}_t)$ for some $\epsilon_t > 0$ and $c_t \triangleq f_0(\bar{q}_t) - \min f_0 - \epsilon_t$. Since $\eta' + \gamma \geq$

$f_0(\hat{y}) - \min f_0$ by Corollary 2.4 (applied similarly to Corollary 2.5 above), we may approximate

$$\begin{aligned}\eta_t \triangleq \eta' + \gamma - c_t - \epsilon_t &\geq f_0(\hat{y}) - \min f_0 - c_t - \epsilon_t \\ &= f_0(\hat{y}) - f_0(\bar{q}_t) + (f_0(\bar{q}_t) - \min f_0 - \epsilon_t) - c_t \qquad (7.14) \\ &\geq f^t_{\mathrm{TSP}}(\hat{y}) - f^t_{\mathrm{TSP}}(\bar{q}_t) \quad \text{for } \hat{y} \in E_t \cap D.\end{aligned}$$

The last inequality follows since $f_0 \leq f^t_{\mathrm{TSP}}$ with equality on E_t. The inclusion in claim (iii) now follows from Lemma 2.4 applied to $f = f^t_{\mathrm{TSP}}$ and $\nu = 0$ with $\epsilon' = 0$ and $\overline{\epsilon} > \epsilon_t$, since then $\hat{y} \in U_{\bar{q}_t}(\eta_t, 0) \subset \bar{q}_t + \bigcap_{\epsilon \in (0,\overline{\epsilon})} (\eta_t + \epsilon)(\partial_\epsilon f^t_{\mathrm{TSP}}(\bar{q}_t))^\circ \subset \bar{q}_t + (\eta' + \gamma - c_t)C_t$. We have taken $\epsilon = \epsilon_t$ for the last inclusion.

Finally, for claim (iv), note that since $f^t_{\mathrm{TSP}}(\hat{y}) \geq f^t_{\mathrm{TSP}}(\bar{q}_t)$, (7.14) implies $\eta' + \gamma - c_t \geq \epsilon_t > 0$. Now just expand c_t in this condition. □

Suppose that $\hat{y}$ is an (approximate) minimiser of the perturbed problem $f^\lambda_{\mathrm{KMTSP}}$ in a predetermined neighbourhood D of any $\bar{a}_{\hat{\sigma}}$. The first two claims of Theorem 7.3 then say that for small $\lambda > \lambda_0 = 1/2$, $\hat{y}$ actually belongs to a smaller set that behaves quite well with respect to λ and $\epsilon \geq 0$. The fourth claim says that for $\hat{y}$ to not belong to the predetermined neighbourhood of some $\bar{a}_{\hat{\sigma}}$, λ or ϵ must be large enough (since $f_0(\bar{q}_t) > \min f_0$ for $\bar{q}_t \neq \bar{a}_{\hat{\sigma}}$). Therefore, for small enough $\lambda > 1/2$, the minimisers of $f^\lambda_{\mathrm{KMTSP}}$ stay within a linearly-scaled region around $\bar{a}_{\hat{\sigma}}$.

While the optimal solution appears in the local bound in $C^\circ_{\hat{\sigma}}$, applying the argument proving its boundedness in the proof, we can approximate it by considering all the possible non-degenerate angles between the points a_k, and choosing the smallest ones. That will, of course, increase the bound. Computing the global bound is much more complicated.

Note that claim (i) of Theorem 7.3 provides a necessary condition for a local minimisers (or, in fact, any point for either the KM or MO reformulation) to be close to a real solution of the Euclidean TSP: if the point $\hat{y}$ can be unambiguously morphed into $\bar{a}_\sigma$ for some, not necessarily optimal σ – which is the case, e.g., when $\hat{y} \in \prod_{i=1}^n \mathbb{B}(a_{\sigma i}, \delta_{\sigma i})$ – the condition (7.13) must hold for the permutation σ to be an optimal path. It suffices to take $\eta = f_{\mathrm{TSP}}(\bar{a}_\sigma) - f_{\mathrm{TSP}}(\hat{y})$ and $D = \{\bar{a}_\sigma, \hat{y}\}$, for if η becomes negative this way, we know a better minimiser and test point.

It remains to discuss the sensitivity in λ of solutions to $f^\lambda_{\mathrm{MOTSP}}$. Clearly, we could still directly apply Corollary 2.5, and may actually show that $0 \in \operatorname{int} C_\epsilon(\bar{a}_{\hat{\sigma}})$ (in the notation of Chapter 2) for $\epsilon = 0$ and then for $\epsilon \in [0, \overline{\epsilon})$ when $\overline{\epsilon} > 0$ is small enough. We could thus approximate $U_{\bar{a}_{\hat{\sigma}}}(\eta, \epsilon') \subset \bar{a}_{\hat{\sigma}} + \eta C^\circ_{\epsilon_\nu}(\bar{a}_{\hat{\sigma}})$ with the polar bounded, by choosing $\epsilon = \epsilon_\nu \leq \epsilon' < \overline{\epsilon}$, the latter two values to be determined.

Alternatively, by Lemma 7.3 and (7.12), we could for $\lambda_1 \in (\lambda_0, 1/2)$ approximate $f^{\lambda_0}_{\mathrm{MOTSP}}(\hat{y}) - f^{\lambda_0}_{\mathrm{MOTSP}}(\bar{a}_{\hat{\sigma}})$ from below in $E^{\lambda_1}_\sigma$ with

$$\begin{aligned}f^{\lambda_0}_{\mathrm{MOTSP}}(\hat{y}) - f^{\lambda_0}_{\mathrm{MOTSP}}(\bar{a}_{\hat{\sigma}}) &= f_{\mathrm{MO}}(\hat{y}; \bar{a}) - f_{\mathrm{MO}}(\bar{a}_{\hat{\sigma}}; \bar{a}) - \lambda_0(f_{\mathrm{TSP}}(\bar{a}_{\hat{\sigma}}) - f_{\mathrm{TSP}}(\hat{y})) \\ &\geq (\lambda_1 - \lambda_0) r_{\hat{\sigma}}(\hat{y}) = \max(y_i - a_{\hat{\sigma} i})^T C_{\hat{\sigma}},\end{aligned}$$

where $C_{\hat{\sigma}} \triangleq 2(\lambda_1 - \lambda_0) \prod_{i=1}^n \mathbb{B}(0,1)$. Then we could apply the gauge-inversion arguments in Section 2.4.1 to get the obvious bound. However, as the neigh-

bourhoods $E_\sigma^{\lambda_1}$ are merely proved to exist (when $m > 1$, and $\lambda_1 > 0$, which we require), the bounds so obtained would be rather poor compared to $f_{\mathrm{KMTSP}}^{\lambda}$.

7.5 Heuristics

As we shall see in Section 7.6, the performance of our basic algorithm is not all that great for larger instances of the Euclidean TSP. Therefore, in this section, we consider various heuristic approaches that could be used to speed up the algorithm or improve the results otherwise. As a first task, however, the association heuristic demands some clarification.

7.5.1 The association heuristic

The proof of Theorem 7.1 provides a conceptual algorithm for obtaining a permutation σ from any sequence of points $\bar{y} = (y_1, \ldots, y_n)$:

1. Assign the points a_k to the closest y_j, forming the cluster C_j (handling ambiguous cases arbitrarily).
2. Remove all the points y_j with empty clusters.
3. Re-insert points in the path, at any $a_k \in C_j$, $a_k \neq y_j$ (the closest in our implementations), before or after y_j (depending on which seems to provide shorter path).
4. Repeat steps 1–3 while there is something to be done.

Note that when C_j consists of a_k alone (and there were no ambiguous assignments), these steps amount to moving y_j at a_k, as y_j would be removed after the new point has been placed at a_k.

Any reinsertion may change the clusters, the new (reinserted) point assimilating points from clusters of y_i, for $i \neq j$ as well as j. If we ignore this fact for $i \neq j$, we may construct σ locally in a hierarchic fashion, "splitting" each cluster until it consist of a single a_k. Otherwise we need to recalculate/shuffle the clusters after each reinsertion. Some improvements to the resulting path length can sometimes be obtained this way, but the method is quite dependent on the order of processing.

7.5.2 Number of cluster centres

A straightforward heuristic derived from our reformulations in the earlier sections, is to reduce the number K of the points y_i used in the minimisation method. After the "shape" of the path has been obtained with a reduced number of points, it can then be refined by adding more points using the already described rules for associating (unassociated and duplicate) points with cities. In case of the MO variant, when $K < n$, we have to alter the factor of the function ν_{MO}, in order

to keep the objective function level-bounded, and for reasonable results. Our somewhat arbitrary but obvious choice of factor is $n/(2K)$, which is below the $n/(2K-2)$ upper bound from Chapter 6:

$$\nu_{\mathrm{MO}}^{K}(\bar{y}) \triangleq \frac{n}{2K} \sum_{i=1}^{K} \sum_{j=1}^{K} \|y_i - y_j\|.$$

Notice that the upper bound for λ ensuring that $\lambda f_{\mathrm{TSP}} - \nu_{\mathrm{MO}}^{K}$ is concave, increases similarly, and we have indeed used $\lambda = n/K$ in our experiments.

7.5.3 Hierarchical clustering.

An obvious refinement of the previous heuristic is analogous to hierarchical clustering:

1. Run our path-length perturbed clustering algorithm on the whole data, with a small number K of clusters.
2. Assign each a_k to the closest y_i, producing the cluster C_i.
3. Run the algorithm again on C_i with a new set of "cluster centres", of size $K_i \leq \#C_i$. Continue this subdivision until the size of the cluster C_i is small enough to merit choosing $K_i = \#C_i$.
4. Construct the full path by combining the paths of the lowest-level clusters along the paths formed by the higher-level clusters centres.

There is a small problem with this approach as such: the paths are closed, so combining them will produce unnecessary detours. However, this is no big problem: we just have to alter f_{TSP} to not attract the first and last points of the open path we want. We can do more: we can attract the endpoints to points in the previous cluster:

$$f_{\mathrm{TSP}}^{\mathrm{open}}(\bar{y}; a_{\mathrm{prev}}, a_{\mathrm{next}}) \triangleq \|y_1 - a_{\mathrm{prev}}\| + \|y_K - a_{\mathrm{next}}\| + \sum_{j=1}^{K-1} \|y_j - y_{j+1}\|. \qquad (7.15)$$

There are various potential choices for a_{prev} and a_{next}. One is the points y_{i-1} and y_{i+1} in the higher-level path (when we are working on C_i). Another would be the points $a_{\mathrm{prev}} \in C_{i-1}$ and $a_{\mathrm{next}} \in C_{i+1}$ that minimise the distance to C_i. In the experiments of Section 7.6 we have chosen the former. Based on a limited number of tests, the latter more complex approach does not seem to improve the results. Note that the first two terms of $f_{\mathrm{TSP}}^{\mathrm{open}}$ are Euclidean distances from fixed points. They can therefore be included in the convex part of the objective function, when we choose to minimise it with the perturbed Weiszfeld method.

A few more choices remain in the hierarchical algorithm: At which point to run the association heuristic: for the whole path, or for the lowest-level clusters? In the experiments to follow, we have chosen the former combined with the local association heuristic, as this seems to provide the best ratio of time spent

to quality of results. Another available choice is the number of points K_i to use in each cluster. Our somewhat arbitrary choice has been to specify a maximum number $M \geq 2$, but instead of greedily choosing $K_i = \min\{M, \#C_i\}$, we try to do this bottom-up: we try to predictively assign the largest number of points to the lowest-level clusters, by choosing

$$K_i = \lceil \#C_i / M^{\lfloor \log_M \#C_i \rfloor} \rceil \quad \text{when} \quad \#C_i > M. \tag{7.16}$$

This appears to provide better results than the greedy approach, based on a limited number of tests.

7.5.4 Clustering for initial iterate

This approach consists of running the previous heuristic without the association step to obtain an initial iterate for the basic algorithm, that we perform only a few steps of.

7.5.5 Path-following

Yet another approach would be to calculate an approximate solution to a penalised version of the problem for some λ, and then with a smaller one starting from the previous result. Unfortunately, at least with the limitation $\lambda \leq 1$ inherent in the perturbed Weiszfeld method, this does not appear to provide considerably improved results.

7.6 Experiments

We have implemented our algorithms [in Haskell; see Peyton Jones et al., 2003] and tested our method on some problems from TSPLIB [Reinelt, 1991], on an Athlon64 3200+ tabletop computer. In each case, we have used the step size $\omega = 1.4$ in the perturbed Weiszfeld algorithm of Chapter 5: of the values we've tried, it seems to provide the most consistently best results, largely in agreement with experimental results for the plain Weiszfeld algorithm [Äyrämö, 2006, Appendices 2–3]. Although each $\omega \in [1, 2)$ does provably provide a descending sequence of iterates, it would be possible to do a line search step in the algorithm as well. The initial iterate has likewise in each case been with the cities equally distributed on a circle, centred and scaled to fit in the problem data. Such a choice seems to provide generally better results than a (totally) random initial iterate, which may contain self-crossings of the path on a large scale, that our method seems poor at removing.

7.6.1 The basic algorithm

In summary, the basic algorithm consists of

1. Choose an initial iterate $\bar{y}_{[0]}$, step length $\omega \in [1, 2)$, penalty parameter $\lambda \in (0, 1]$, and maximum iterations count or other stopping criterion.

2. Apply the perturbed Weiszfeld method to problem (7.5), to get $\hat{y}$.

3. Use the association heuristic to find a permuted path $\bar{a}_\sigma$ from $\hat{y}$.

The results for this method may be found in Tables 7.1 through 7.4. Furthermore, Figure 7.1 shows results for some simple instances from the first series of tests. In most of the test cases, we have used $\lambda = 1.0$, as it is the upper limit at which the TSP penalty term can certainly be "absorbed" into the concave part of the diff-convex objective, and thus that our algorithm can handle. Lower values also do not appear to provide better results. In each of these tests of the basic algorithm, we have used the "semi-global" variant of this association heuristic discussed in Section 7.5, to obtain a permutation of the points $a_1, \ldots, a_n$ from the results of the Weiszfeld algorithm; cf. Figure 7.1(c). In two problems, Eil101′ and PR1002′, some of the parameters have been varied to offer points of comparison: the problem Eil101′ uses $\lambda = 2.0$, although our algorithm is not entirely applicable for such a choice. In the problem PR1002′ we have used only $K = 50$ y_is in the perturbed Weiszfeld method and added the rest later, as again discussed in Section 7.5.

In the first series, in Table 7.1, the maximum number of iterations of the perturbed Weiszfeld method has been 1000, and the stopping threshold τ (maximum difference in norm between successive iterates) has been 10^{-5}, whereas in Table 7.2 the values have been 10000 and 10^{-2}, respectively. In the third series in Table 7.1, where we have excluded the cases from the second series that used the maximum number of iterations, the values are 10000 and 10^{-5}, respectively. Finally, in Table 7.4, we have allowed for just $10 \log_2 n$ iterations.

In the tables, the "Weiszfeld time" field is the time (in seconds) it took for the perturbed Weiszfeld method to finish, and the field "Weiszfeld iterations" is the number of iterations of this method used. The "Total time" field indicates the time it took in addition to this, to move the resulting points towards the cities, as described above. Such an intermediate result is included in Figure 7.1(c) for the Berlin52 problem. Note that the "TSPLIB path length" is calculated with the Euclidean metric rounded to nearest integer, instead of the plain Euclidean metric, with which "Our path length" has been calculated. Finally, the instance size (n) is indicated by the TSPLIB problem name itself.

As we can see, the results are not all that great, compared to what is achievable with other methods; cf. Johnson and McGeoch [2002, 1997]. Some of the run-time can be attributed to our choice of language: Haskell and the compilers available for it, with standard unoptimised data structures, are not presently quite up to par with lower-level languages in speed, but offer much comfort of implementation. As for the quality of the paths, it can clearly be seen that the relative quality of the results degrades as the number of cities grows. Looking at the figures, our results seem to share a lot of the overall structure of the optimal results, however, which would indicate that they could serve as starting points

TABLE 7.1 Results for max_iters $= 1000$, $\tau = 10^{-5}$, and $\omega = 1.4$

Problem	Berlin52	Eil101	TS225	PR1002	Eil101′	PR1002′
K	n	n	n	n	n	50
λ	1.0	1.0	1.0	1.0	2.0	1.0
Weiszfeld iterations	1000	1000	1000	1000	1000	1000
Weiszfeld time	3.0	10.9	53.1	1315.0	11.5	46.4
Total time	3.1	11.3	56.0	1604.0	11.9	472.9
TSPLIB path length	7542	629	126643	259045	629	259045
Result path length	8951.6	726.0	207730.3	370184.2	706.8	375395.6

TABLE 7.2 Results for max_iters $= 10000$, $\tau = 10^{-2}$, and $\omega = 1.4$

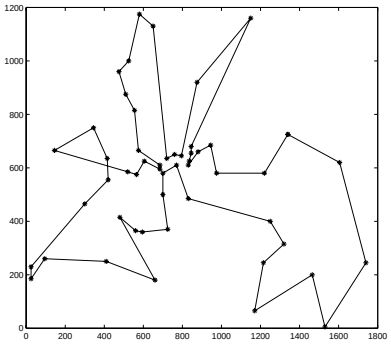
Problem	Berlin52	Eil101	TS225	PR1002	Eil101′	PR1002′
K	n	n	n	n	n	50
λ	1.0	1.0	1.0	1.0	2.0	1.0
Weiszfeld iterations	396	201	797	10000	10000	1875
Weiszfeld time	1.2	2.3	42.6	12877.3	114.1	88.2
Total time	1.3	2.6	45.5	13146.1	114.5	518.7
TSPLIB path length	7542	629	126643	259045	629	259045
Result path length	8951.6	719.7	207730.3	363456.1	702.9	365239.2

TABLE 7.3 Results for max_iters $= 10000$, $\tau = 10^{-5}$, and $\omega = 1.4$

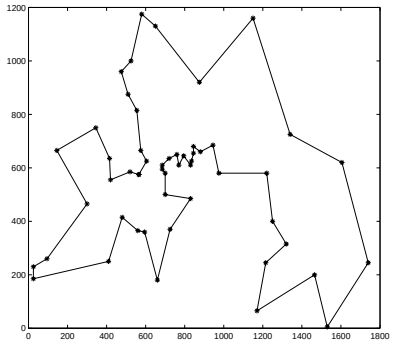
Problem	Berlin52	Eil101	TS225	PR1002′
K	n	n	n	50
λ	1.0	1.0	1.0	1.0
Weiszfeld iterations	1531	2573	2376	2344
Weiszfeld time	4.7	28.3	125.0	112.9
Total time	4.7	28.6	127.8	554.12
TSPLIB path length	7542	629	126643	259045
Result path length	8951.6	706.7	207730.3	365239.2

TABLE 7.4 Results for max_iters $= 10 \log_2 n$, $\tau = 10^{-5}$, and $\omega = 1.4$

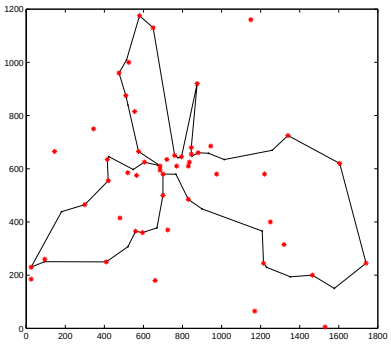
Problem	Berlin52	Eil101	TS225	PR1002	Eil101′	PR1002′
K	n	n	n	n	n	50
λ	1.0	1.0	1.0	1.0	2.0	1.0
Weiszfeld iterations	57	67	78	100	67	100
Weiszfeld time	0.18	0.8	4.6	145.5	0.8	4.8
Total time	0.25	1.2	9.6	617.4	1.2	442.5
TSPLIB path length	7542	629	126643	259045	629	259045
Result path length	9087.1	741.9	210694.5	392377.7	741.9	372602.0



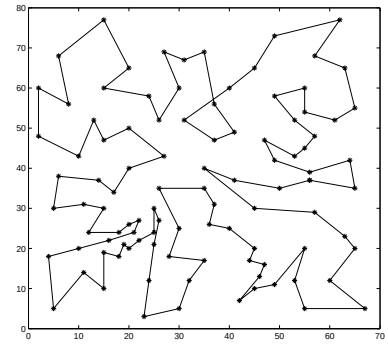
(a) Berlin52: Result, $\lambda = 1.0$



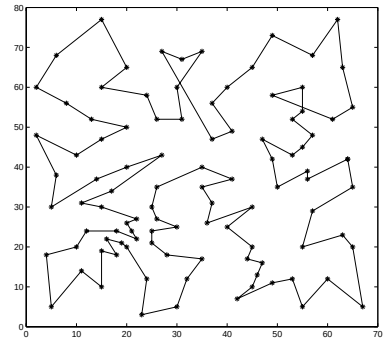
(b) Berlin52: TSPLIB optimal path



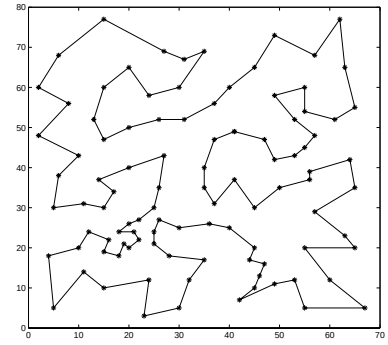
(c) Berlin52: raw Weiszfeld, $\lambda = 1.0$



(d) Eil101: Result, $\lambda = 1.0$



(e) Eil101: Result, $\lambda = 2.0$



(f) Eil101: TSPLIB optimal path

FIGURE 7.1 Results for Berlin52 and Eil101 from TSPLIB

for other methods. Also note comparing Figure 7.1(a) to (c), that our naïve association heuristics induce some clear mistakes, such as self-crossings of the path.

That increasing λ for Eil101' improves the path, seems to be a general trend, although occasionally worse results are obtained as for $\lambda = 1.0$. Decreasing λ below 1.0 also usually seems to degrade the result, as would increasing it too much. In the second series of tests, we also see that the algorithm indeed does not appear to converge.

In the last series with just $10 \log_2 n$ iterations, the performance does not actually decrease relatively that much from series with more iterations. In this series, the result for PR1002 with 50 y_is actually beats the one for all 1002 y_is in the Weiszfeld method. In both cases, considerable time is spent in the (quite unoptimised and naïvely implemented) association heuristic.

7.6.2 The hierarchical algorithm

In summary, this heuristic consists of the steps

1. Choose maximum prototype count $M \geq 2$, as well as parameters for the perturbed Weiszfeld method, and initialise the initial cluster $C_0 = \{a_1, \ldots, a_n\}$.
2. Calculating new prototype count K_i for each cluster from (7.16).
3. Apply the Weiszfeld method with the modified penalty (7.15) and K_i new prototypes on each present cluster C_i. The points a_{prev} and a_{next} are the prototypes of the present prototypes with next and previous index.
4. Split clusters that did not yet have equally many new prototypes and vertices, by the new prototypes. Recursively continue from Step 2.
5. Apply the association heuristic on each completed cluster, and join the in-cluster paths in the order given by the higher-level clusters.

Table 7.5 lists results for this approach. The number "Total Weiszfeld its." in the table, is the total number of iterations of the Weiszfeld algorithm at all scales. As already mentioned in Section 7.5, we have used the local variant of the association heuristic on the full resulting hierarchical Weiszfeld path, to obtain the final permutation.

In this series of experiments, we have used bigger problem instances than in the previous experiments. As can be seen from the results, with this heuristic, the running time becomes noticeably more feasible than that of the basic algorithm, and without degrading the results – improving them, in fact. (For the smallest instances from the other experiments, the heuristic degrades the results, however.) Note that for the biggest instances we only have bounds on the optimal path length from TSPLIB, and for PLA33810 this is, in fact, for the ceiling of the Euclidean distance, instead of rounded.

Using only a small number of iterations has been more our goal in this series of tests than obtaining the best possible result we can with our algorithms. By

TABLE 7.5 Clustering heuristic results (max_iters $= 10 \log_2 \#C$, $\tau = 10^{-5}$, $\omega = 1.4$)

Problem	PR1002			PR2392		
M	50	100	150	50	100	150
Total Weiszfeld its.	1741	1033	835	4472	2606	1743
Weiszfeld time	5.2	7.7	10.6	27.2	24.5	32.4
Total time	5.4	8.0	10.9	28.6	25.8	33.7
TSPLIB path length	259045			378032		
Result path length	345380	346628	346902	558451	535006	521040

Problem	RL11849			PLA33810		
M	50	100	150	50	100	150
Total Weiszfeld its.	21298	12759	8980	62609	34811	24940
Weiszfeld time	122.3	288.4	454.1	586.9	1270.3	2067.3
Total time	172.2	339.0	504.5	1037.0	1720.6	2518.7
TSPLIB path l. bnd.	[920 847, 923 368]			[65 913 275, 66 116 530]		
Result path length	1 410 087	1 386 317	1 360 373	99 304 887	97 915 373	96 554 643

TABLE 7.6 Average results for random instances

Algorithm	Basic		Clustering 50,100 & 150	
Problem set	uniform 1k-3k	clustered 1k-10k	uniform 1k-3k	clustered 1k-10k
# samples	15	15	54	54
Average performance	1.66	1.49	1.42	1.36

using two times as many steps in each cluster ($20 \log_2 \#C$), we could still improve some of the results noticeably, whereas others would simply take longer to compute without much improvement. (More meticulous choice of τ could of course be used to control the number of steps as well.) Likewise, using the hierarchical method with a small number of iterations of the basic Weiszfeld method to obtain an initial iterate, as discussed in Section 7.5, would slightly improve the results. For larger instances there would be a noticeable increase in time spent, however.

Notice, nevertheless, that the results appear to fall approximately around 1.5 times the optimal path length (modulo slightly differing distance measures). Further evidence for this is provided in Table 7.6. There, we have calculated the average performance of our methods for the 1-3k city random and 1-10k city random clustered Euclidean instances of the TSP DIMACS challenge problems from Johnson and McGeoch [2002].[2] The average for the clustering heuristic is further taken over all the parameter values $M = 50, 100, 150$. The performance reported is the proportion of the path length calculated by our algorithm, to the Held-Karp bound for the problem. Our methods appear to perform better for the clustered than non-clustered instances, as can be expected.

7.6.3 Use as an initial tour

We also tested in a few cases, the use of our method for providing an initial tour for other methods: LKH [Helsgaun, 2000], Concorde [Applegate et al., 1998], and basic 2-Opt. All of these methods improved upon the initial tour from our

[2] For the data, see `http://www.research.att.com/~dsj/chtsp/`.

method. Unfortunately, our method did not significantly improve upon a random or default initial tour: LKH and Concorde did in fact seem to take longer in their computations. The 2-Opt results varied, with the initial tour from our method occasionally providing significant improvements in the final results, and at other times slightly worse results. (The results obviously depend on the processing order in the implementation of the method.) It seems to us that these non-geometrical algorithms fail to exploit the overall shape of the path that our method seems to approximate, with the errors being mostly (but not exclusively) on the small scale.

8 CONCLUSIONS

We have provided both new general theoretical results for diff-convex functions, as well as more applied mathematical results in relation to some location problems. General mathematical results were provided on optimality and sensitivity of diff-convex functions, along with a characterisation of level-boundedness. We also studied the internal structure of a special form of diff-convex functions, and based on that study, proved local convergence for an extension of interior point methods for linear programs on symmetric cones. A globalisation strategy was also provided based on the idea of the filter method. The resulting method was shown to converge polynomially in r to ϵ-semi-critical points under conditions related to the degree of level-boundedness and reinitialisation quality. We also extended the Weiszfeld method to problems of "perturbed spatial medians" and proved its convergence to semi-critical points under some constraints.

On the application side, we provided a new clustering formulation, and theoretically studied the applicability of the above-mentioned methods – the Weiszfeld method in particular – to this problem, as well as the classical multisource Weber problem. We then showed a relationship of these problems to the Euclidean TSP, and again studied the application of the above-mentioned optimisation methods.

While our focus was theoretical, some numerical results were also provided. Although the performance of the interior point methods was discussed briefly, we concentrated on the Weiszfeld method, as it seemed to be more promising in practise – and demanding of far less meticulous parameter-tuning. The experiments for the clustering problems were concentrated on comparing the two objectives. More study remains for more practically oriented researchers, in particular in relation to the performance of our methods in comparison to other methods, such as the classical K-means -style method. In case of the Euclidean TSP, our tests were slightly more extensive. While the Weiszfeld method managed to produce rather reasonably-shaped paths in a low number of iterations, the results were not all that good in comparison to existing methods for the TSP. Nevertheless, our results could perhaps be improved upon, by using other optimisations methods, parameters, and heuristics.

APPENDIX 1 LOCAL MINIMA OF K-MEANS TYPE PROBLEMS

Consider the K-means-type problem

$$\min_{\bar{w},\bar{y}} \left(f(\bar{y};\bar{w}) \triangleq \sum_{i=1}^{n}\sum_{j=1}^{K} w_{ij} d(a_i, y_j) \right) \text{ with } w_{ij} \in \{0,1\} \text{ and } \sum_{j=1}^{K} w_{ij} = 1. \quad \text{(A1.1)}$$

Here d are some distance functions and the data $\{a_i\}_{i=1}^n \subset \mathbb{R}^m$. The weights w_{ij} indicate to which cluster j the vertex i belongs to, and $\bar{y} = (y_1, \ldots, y_K) \in \mathbb{R}^{mK}$ are the cluster prototypes.

The K-means-type algorithm for (A1.1) assigns each a_i to the closest prototype y_j (setting $w_{ij} = 1$), and calculates new prototypes y'_j by minimising $\sum_{i=1}^n w_{ij} d(a_i, y'_j)$. This procedure is then repeated until there is no change in the assignments.

Selim and Ismail [1984] prove the convergence of this method to (differentiable) Karush-Kuhn-Tucker (KKT) points of the objective function under the relaxed constraint $w_{ij} \in [0,1]$, which does not affect the optima of (A1.1). They also make claims on local optimality of these points. However, these latter results are not entirely correct, as their characterisation of local optimality by directional derivatives [Selim and Ismail, 1984, Lemma 7], quoted without proof, is incorrect for non-convex functions. Directional derivatives being non-negative to all feasible directions is not sufficient, merely necessary, for local optimality of non-convex functions. However, positivity of the directional derivatives is sufficient.[1] In this appendix, we provide corrections to the results depending on this incorrect characterisation.

We define the reduced objective function as $F(\bar{w}) \triangleq \min_{\bar{y}\in\mathbb{R}^{mK}} f(\bar{y};\bar{w})$, and the feasible polytope of weights as $\mathcal{W} \triangleq \{\bar{w} \mid \sum_j w_{ij} = 1, w_{ij} \geq 0\}$. Problem (A1.1) may then be recast as

$$\min_{\bar{w}\in\mathcal{W}} F(\bar{w}). \quad \text{(A1.2)}$$

We denote the set of bounded minimisers for a weight $\bar{w}$ as $P(\bar{w}) \triangleq \{\bar{y} \in V \mid \bar{y} \text{ minimises } f(\bar{y};\bar{w})\}$, and set $J_i \triangleq J_i(\bar{y}) \triangleq \{j \mid j \text{ minimises } d(a_i, y_j)\}$.

We require that the minimum of $f(\cdot;\bar{w})$ is reached in some compact set V for every $\bar{w}$, whence in fact $F(\bar{w}) = \min_{\bar{y}\in V} f(\bar{y};\bar{w})$. This is a valid assumption for most distances of interest, as shown by the following lemma:

---

[1] This can be seen from the equivalence of the normal directional derivative to the Hadamard lower directional derivative

$$\underline{d}F(\bar{w};z) \triangleq \liminf_{t\searrow 0, v\to z} \frac{F(\bar{w}+tv) - F(\bar{w})}{t}$$

for finite concave functions [cf., e.g., Penot 1978]. For, if $\underline{d}F(\bar{w};\cdot) > 0$, and $\bar{w}_{[k]} \to \bar{w}$ with $F(\bar{w}_{[k]}) < F(\bar{w})$, we get a contradiction by setting $v_{[k]} \triangleq (\bar{w}_{[k]} - \bar{w})/t_{[k]}$ and $t_{[k]} \triangleq \|\bar{w}_{[k]} - \bar{w}\|$, and choosing a subsequence of $v_{[k]}$ convergent to some feasible direction z. This can be done in our finite-dimensional setting. Then $\underline{d}F(\bar{w};z) \leq \lim_{k\to\infty}(F(\bar{w}_{[k]}) - F(\bar{w}))/t_{[k]} \leq 0$.

Lemma A1.1. *Suppose each $d(a_i, \cdot)$ is a finite convex function minorised by $c\|\cdot\| - b$ for some norm $\|\cdot\|$ and constants $c > 0$, $b \geq 0$. If also $\bigcap_{i=1}^n \operatorname{dom} d(a_i, \cdot) \neq \emptyset$, then $f(\cdot; \bar{w})$ reaches its minimum in some compact set V for every $\bar{w} \in \mathcal{W}$.*

Proof. That $f(\cdot; \bar{w})$ has minimisers in a compact set V follows if $f_j : y_j \mapsto \sum_{i=1}^n w_{ij} d(a_i, y_j)$ always has minimisers in a compact set V'. It thus suffices to consider the single-facility case.

When $r \triangleq \sum_{i=1}^n w_{ij} = 0$, the minimising y_j may be taken in any compact set of choice. If $r > 0$, then for a minimising y_j

$$\max_i d(a_i, 0) \geq f_j(0)/r \geq f_j(y_j)/r \geq \sum_{i=1}^n c(w_{ij}/r)(\|y_j\| - b) = c\|y_j\| - cb.$$

Thus a large y_j cannot minimise f_j. □

The following theorem sharpens and fixes [Selim and Ismail, 1984, Theorem 8] along with providing a condition of non-optimality for cases excluded by these weakened claims. Some differentiability assumptions are made in the claim, because in cases with only subdifferentiability present, small perturbations in weights do not necessarily disturb optimality of individual prototypes.

Theorem A1.1. *Suppose each $d(a_i, \cdot)$ is a finite convex function, and that $f(\cdot; \bar{w})$ has finite minimum in some compact set V for all $\bar{w}$. Let $\bar{w}^* \triangleq \{w_{ij}^*\}$ be an extreme point of $\mathcal{W}$. Then $\bar{w}^*$ is a local minimiser of the problem* (A1.2) *if for all $\bar{y}^* \in P(\bar{w}^*)$ and $i = 1, \ldots, n$, we have $\#J_i(\bar{y}^*) = 1$, and*

$$F(\bar{w}^*) = f(\bar{y}^*; \bar{w}^*) \leq \min_{\bar{w} \in \mathcal{W}} f(\bar{y}^*; \bar{w}). \tag{A1.3}$$

The allocation $\bar{w}^*$ is not a local minimiser if (i) the condition (A1.3) *does not hold, or (ii) if $\#J_{i'}(\bar{y}^*) > 1$ for some $i' \in \{1, \ldots, n\}$ and for some $j' \in J_{i'}(\bar{y}^*)$, (a) $\nabla d(a_{i'}, \cdot)(y_{j'}) \neq 0$, and (b) $d(a_i, \cdot)$ is differentiable at $y_{j'}$ for all i with $j' \in J_i(\bar{y})$.*

Proof. The function F is concave [cf. Selim and Ismail, 1984]. As discussed above, for $\bar{w}^*$ to be a local minimiser, it is sufficient that $F'(\bar{w}^*; z) > 0$ for all feasible directions z. By [Danskin, 1966, Theorem 1], utilising the compactness of V,

$$F'(\bar{w}^*; z) = \min\{\nabla_{\bar{w}^*} f(\bar{y}; \bar{w}^*)^T z \mid \bar{y} \in P(\bar{w}^*)\} = \min\{f(\bar{y}; z) \mid \bar{y} \in P(\bar{w}^*)\}.$$

For extremal $\bar{w}^*$, the feasible directions transfer weight from assignments with $w_{ij} = 1$ to assignments $w_{ij'} = 0$ for $j' \neq j$. If (A1.3) holds, it follows from $\#J_i(\bar{y}^*) = 1$ that the value of $f(\bar{y}^*; \cdot)$ increases by such change for all $\bar{y}^* \in P(\bar{w}^*)$. Therefore $f(\bar{y}^*; z) > 0$, and by the compactness of $P(\bar{w}^*)$, $F'(\bar{w}^*; z) > 0$.

The necessity of (A1.3) for local optimality is immediate from the linearity of f in $\bar{w}$.

Suppose then that (A1.3) holds. If $\#J_{i'}(\bar{y}) > 1$, we still have $f(\bar{y}^*; z) \geq 0$ and thus $F'(\bar{w}^*; z) \geq 0$. Any z that shifts weight between j' and some $j \in J_{i'}(\bar{y})$, $j \neq j'$, will not change the value of f. However, the optimality condition $0 \in \partial\big(\sum_i w_{ij'} d(a_i, \cdot)\big)(y_{j'})$ for the prototype $y_{j'}$ under fixed $\bar{w}$, will be upset under the assumptions (a) and (b). Hence the value of f can be improved from $F(\bar{w}^*; \bar{y}^*)$ by altering $\bar{y}^*$. In consequence $\bar{w}^*$ cannot be a local minimiser. □

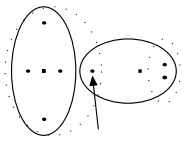


FIGURE A1.1 KKT point with two non-collinear clusters, but not local optimum. The arrow points to the "disputed" vertex and the dotted lines indicate the optimal clusters.

Corollary A1.1. *If $(\bar{w}^*, \bar{y}^*)$ is a KKT point of f, $P(\bar{w}^*)$ is a singleton, and J_i is singleton for each $i = 1, \ldots, n$, then $\bar{w}^*$ is a local minimiser of F.*

Proof. See [Selim and Ismail, 1984, Theorem 9]. □

As we have seen, the results of Selim and Ismail [1984] do not hold if two prototypes y_j and y_k are at the same minimal distance from a vertex a_i under some additional conditions on the distances of these points to other vertices in their corresponding clusters.

In particular, let $d(a, y) = \|a - y\|_2^2$ be the squared distance employed in the *K*-means. We then have everywhere differentiability, and the additional non-optimality conditions (a) and (b) reduce to $d(a_{i'}, y_j) = d(a_{i'}, y_{j'}) > 0$. The case of zero distances obviously depends on the numbers of vertices and prototypes.

In case of the Euclidean metric $d(a, y) = \|a - y\|_2$ or other Minkowski metrics, (a) and (b) reduce to $\min_{i:j' \in J_i} d(a_i, y_j) > 0$, because non-differentiability can happen only when $y_j = a_i$. A sketch of the situation is provided in Figure A1.1. If an undisputed vertex is at zero distance from a prototype, whether we have a minimum is a more complicated issue. It depends on the magnitude of the gradient of the sum of distances to remaining vertices of the cluster, as the subdifferentials in the optimality condition can provide some slack; cf. Chapter 5.

Finally, we note that also in [Selim and Ismail, 1984, Theorem 12], the condition $F(\bar{W}) \leq F(W)$ needs to be changed into strict inequality to reflect the corrected optimality condition.

APPENDIX 2 THE WEISZFELD DIRECTION IN NON-PARTIALLY-OVERLAPPING CASE

As noted in Section 5.2, 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

$$\min_{z \in Z(p)} \left(g^T z + \sum_{k \in \pi} \|z\|_k \right) \tag{A2.1}$$

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 $\rho_k = \mathrm{diag}(1, 1, 0)$, $\rho_i = \mathrm{diag}(0, 1, 0)$, as well as $\rho_k = \mathrm{diag}(1, 0, 0)$, $\rho_k = \mathrm{diag}(0, 0, 1)$, but are not satisfied in cases like $\rho_k = \mathrm{diag}(1, 1, 0)$, $\rho_i = \mathrm{diag}(0, 1, 1)$.

To start solving (A2.1), 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

$$\hat{z} = -\sum_{\rho \in \psi} \beta_\rho g_\rho \tag{A2.2}$$

for some $\beta_\rho \geq 0$ and $g_\rho \triangleq \rho g$; see Valkonen [2006] 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 A2.1. *Suppose ρ_τ is maximal (in $\sqsubset$). If $\hat{\beta}_\tau > 0$ and $\|\rho_\tau \hat{z}\| > 0$, then $\hat{\beta}_\tau \propto 1 - w_\tau/\theta_\tau$ (with respect to scaling of the final result), and $\hat{\beta}_k = \hat{\beta}_\tau \gamma_k$ for $\rho_k \sqsubset_! \rho_\tau$, where*

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

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 (A2.1) is a convex problem, and therefore the Karush-Kuhn-Tucker conditions being fulfilled is sufficient for a minimum. Let $\alpha_k \triangleq \|\rho_k z\| = \|\sum_{\rho' \sqsubset \rho_k} \beta_{\rho'} g_{\rho'}\|$. Then, inserting (A2.2) into (A2.1), differentiating with respect to β_ρ, adding the constraints $-\beta_\rho \leq 0$ and $\|z\|^2 \leq 1$, we get after dividing by $\|g_\rho\|^2$,

$$\lambda_\rho \geq 0,\ \lambda_\rho \beta_\rho = 0 \,\forall \rho \in \psi, \quad \lambda \geq 0,\ \lambda(\|z\|^2 - 1) = 0$$
$$1 - \sum_{k \in \pi : \rho \sqsubset \rho_k} w_k \delta\Big(\frac{\beta_\rho}{\alpha_k}\Big) - \lambda \beta_\rho + \lambda_\rho \ni 0, \forall \rho \in \psi, \tag{A2.3}$$

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 (A2.3) 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 (A2.3) 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 $\alpha'_k \triangleq \alpha_k/\hat{\beta}_\tau = \|\sum_{\hat{\rho}_j \sqsubset \rho_k} \gamma_j g_{\hat{\rho}_j}\|$. 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 A2.1. *Lemma A2.1 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 (A2.3) 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 A2.1. Theorem 5.2 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 5.3 may therefore be applied to this sub-problem 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.

APPENDIX 3 LEMMAS ON SUBDIFFERENTIALS

In this section we include a few simple results on convex (approximate) subdifferentials needed in the sensitivity analysis of Section 7.4, that do not seem to appear in the standard literature. First we have the rather obvious,

Lemma A3.1. *Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ be convex, closed, proper, and level-bounded. Then $0 \in \operatorname{int} \mathcal{R}(\partial f)$.*

Proof. Since f is proper, lower-semicontinuous, and level-bounded, it has a finite minimum. We may assume without loss of generality, that $0 \in \partial f(0)$. Denote $A \triangleq \operatorname{cl} \mathcal{R}(\partial f)$. The set A is then convex [Rockafellar, 1972, Section 24]. Suppose $0 \in \operatorname{bd} A$. Then there exists a direction $z \in N_A(0)$, the normal cone to A at 0, with $z \neq 0$. Thus in particular $z^T \partial f(\alpha z) \leq 0$ for all $\alpha \geq 0$. But by monotonicity (abusing notation slightly), $(z-0)^T(\partial f(\alpha z) - 0) = z^T \partial f(\alpha z) \geq 0$. Thus $z^T \partial f(\alpha z) = 0$ for all $\alpha \geq 0$. But then $f(0) \geq f(\alpha z) + \partial f(\alpha z)^T(0 - \alpha z) = f(\alpha z)$ for all $\alpha \geq 0$ in contradiction to level-boundedness. □

Lemma A3.2. *Let $f : \mathbb{R}^m \to \mathbb{R}$ be convex, continuous, and level-bounded, achieving its minimum at $\hat{y}$. Then $0 \in \operatorname{int} \partial_\epsilon f(\hat{y})$ for $\epsilon > 0$.*

Proof. By Lemma A3.1, for small $r > 0$, $\mathbb{B}(0, r) \subset \operatorname{int} \mathcal{R}(\partial f)$. Let now

$$\epsilon(r) \triangleq -\min_{y \in \mathbb{R}^m} g(y, r) \triangleq -\min_{y \in \mathbb{R}^m} \big(f(y) - f(\hat{y}) - r\|y - \hat{y}\|\big).$$

The function g is continuous, and since $\operatorname{int} \mathcal{R}(\partial f) \supset \mathbb{B}(0, r) = \mathcal{R}(\partial(r\| \cdot - \hat{y}\|))$ with the latter closed, $g(\cdot, r)$ is level-bounded by Theorem 2.6. Since g is a decreasing function of r, it is also locally uniformly level-bounded. Thus for small $r > 0$, the function ϵ is continuous by [Rockafellar and Wets, 1998, Theorem 1.17] and finite (by the showed properties of g). As, in fact,

$$\epsilon(r) = -\min_{\bar{z} \in \mathbb{B}(0,r)} \min_{y \in \mathbb{R}^m} \big(f(y) - f(\hat{y}) - z^T(y - \hat{y})\big),$$

we have [cf. Hiriart-Urruty and Lemaréchal, 1993, Section XI] that $\partial_{\epsilon(r)} f(\hat{y}) \supset \mathbb{B}(0, r)$. Finally, since ϵ is continuous and increasing with $\epsilon(0) = 0$, we can find for small enough $\epsilon > 0$ an $r(\epsilon) > 0$, such that $\partial_\epsilon f(\hat{y}) \supset \mathbb{B}(0, r(\epsilon))$. From this the claim follows for small ϵ, and then for all from the nesting of the approximate subdifferentials. □

APPENDIX 4 THE EUCLIDEAN STEINER TREE PROBLEM

In the Euclidean Steiner tree problem, given points $a_1, \ldots, a_n$ (with $n \geq 3$ to rule out the trivial case), we are supposed to find a tree structure connecting these points, and possibly some additional points y_i, so as to minimise the sum of the edges of the tree. In essence, one minimises the size of the minimal spanning tree over the graph consisting of the a_k and the additional Steiner points y_i. There are at most $K \triangleq n-2$ of these according to Gilbert and Pollak [1968], which is our general reference for the basic properties of Steiner trees used below. We may then assume that there are exactly K Steiner points, the extra points not affecting the weight of the optimal solution. Let $\mathcal{T}_{2n-2}$ denote the possible tree structures (their edges) on these $2n-2$ points. Actually we would only have to consider a subset of trees, where each node has degree at most 3, because the minimal angle between lines from a vertex in a Steiner tree is 120 degrees. Even with such reductions, the set of trees is still huge, but fortunately, as we will see, we do not have to care about that.

We may write the objective function for finding the extra *nodes* of the minimal Steiner tree as

$$\min_{\bar{p}} \min_{E \in \mathcal{T}_{2n-2}} \sum_{(q,q') \in E} \|q - q'\|.$$

In fact, since in a full tree (with $K = n-2$ Steiner points, $n \geq 3$), no a_k is directly connected to a_ℓ for $k \neq \ell$, we only have to consider trees on the y_i:

$$\min_{\bar{p}} \left(\sum_{k=1}^{n} \min_{j=1,\ldots,K} \|a_k - y_j\| + \min_{E \in \mathcal{T}_K} \sum_{(y_i,y_j) \in E} \|y_i - y_j\| \right). \tag{A4.1}$$

The first term in this expression is of course the familiar K-spatial median objective function, and admits a diff-convex presentation. But so does the latter term, in a similar manner:

$$\min_{E \in \mathcal{T}_K} \sum_{(y_i,y_j) \in E} \|y_i - y_j\| = \sum_{i=1}^{K} \sum_{j=1}^{K} \|y_i - y_j\| - \max_{E \in \mathcal{T}_K} \sum_{(y_i,y_j) \notin E} \|y_i - y_j\|$$

The Euclidean Steiner tree problem can thus be given a formulation that very closely resembles the formulations for the K-spatial median, and the Euclidean travelling salesperson problem: diff-convex with double sums of Euclidean norms and taking maxima in the concave part. Furthermore, the term involving maxima or minima over $\mathcal{T}_K$ is in fact relatively easy to calculate: we only have to find a minimal spanning tree for the already prescribed points $y_1, \ldots, y_K$, and this can be done in $O(n^2)$ time with Prim's algorithm.

The minimal spanning tree may, however, not be unique, and to calculate the full subdifferential, all the minimal spanning trees should be found. In some methods a single solution suffices, however, as only some subgradient is needed. That would be the case with a further generalisation of the perturbed Weiszfeld algorithm, to arbitrary symmetric and positive-semidefinite weight matrices W_k:

$d_k(y) \triangleq \|W_k(y - a_k)\|$, which include expressions of the form $\|y_i - y_j\|$. Unfortunately, while the method remains descending with straightforward generalisation of the expression for the search direction, it does not appear to be convergent then, even to "semi-critical" points. (In particular, Lemma 5.4 does not go through.)

SCP (Section 4.4) or K-means -style local convex optimisation on (A4.1), however works in conjunction with an interior point method; cf. Remark 4.3. The latter approach along with some heuristic improvements are studied by Dreyer and Overton [1998].

REFERENCES

Alizadeh, F. and Goldfarb, D. [2003]. Second-order cone programming. *Mathematical Programming*, 95(1), pp. 3–51.

An, L. T. H. and Tao, P. D. [2005]. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133(1), pp. 23–46.

Andersen, K. D., Christiansen, E., Conn, A. R., and Overton, M. L. [2000]. An efficient primal-dual interior point method for minimizing a sum of Euclidean norms. *SIAM Journal on Scientific Computation*, 22(1), pp. 243–262.

Applegate, D., Bixby, R., Chavátal, V., and Cook, W. [1998]. On the solution of traveling salesman problems. In *Documenta Mathematica Extra Volume ICM 1998 Berlin*, pp. 645–656.

Arora, S. [1998]. Polynomial time approximation schemes for Euclidean traveling salesman and other geometric problems. *Journal of the ACM*, 45(5), pp. 753–782.

Arora, S. [2003]. Approximation schemes for NP-hard geometric optimization problems: a survey. *Mathematical Programming*, 97, pp. 43–69.

Arora, S., Raghavan, P., and Rao, S. [1998]. Approximation schemes for Euclidean k-medians and related problems. In *ACM Symposium on Theory of Computing*, pp. 106–113.

Attouch, H. and Wets, R. J.-B. [1991]. Quantitative stability of variational systems: I. The epigraphical distance. *Transactions of the American Mathematical Society*, 328(2), pp. 695–729.

Attouch, H. and Wets, R. J.-B. [1993]. Quantitative stability of variational systems: II. A framework for nonlinear conditioning. *SIAM Journal on Optimization*, 3(2), pp. 359–381.

Ausiello, G., Crescenzi, P., Gambosi, G., Kann, V., Marchetti-Spaccamela, A., and Protasi, M. [1999]. *Complexity and Approximation: Combinatorial optimization problems and their approximability properties*. Springer-Verlag.

Bentley, J. J. [1992]. Fast algorithms for geometric traveling salesman problems. *ORSA Journal on Computing*, 4(4), pp. 887–411.

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

Bonnans, J. F. and Shapiro, A. [1998]. Optimization problems with perturbations: A guided tour. *SIAM Review*, 40(2), pp. 228–264.

Brimberg, J., Hansen, P., Mladenović, N., and Taillard, E. D. [2000]. Improvements and comparison of heuristics for solving the uncapacitated multisource Weber problem. *Operations Research*, 48(3), pp. 444–460.

Brimberg, J. and Mladenović, N. [1999]. Degeneracy in the multi-source Weber problem. *Mathematical Programming*, 85(1), pp. 213–220.

Buttazzo, G. and Stepanov, E. [2004]. Minimization problems for average distance functionals. *Calculus of Variations: Topics from the Mathematical Heritage of Ennio De Giorgi, D. Pallara (ed.), Quaderni di Matematica, Seconda Università di Napoli, Caserta*, 14, pp. 47–83.

Chen, P.-C., Hansen, P., Jaumard, B., and Tuy, H. [1992]. Weber's problem with attraction and repulsion. *Journal of Regional Science*, 32(4), pp. 467–486.

Chen, P.-C., Hansen, P., Jaumard, B., and Tuy, H. [1998]. Solution of the multi-source Weber and conditional Weber problems by D.-C. programming. *Operations Research*, 46(4), pp. 548–562.

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

Cooper, L. [1964]. Heuristic methods for location-allocation problems. *SIAM Review*, 6(1), pp. 37–53.

Cox, D. R. [1957]. Note on grouping. *Journal of the American Statistical Association*, 52(280), pp. 543–547.

Danskin, J. M. [1966]. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4), pp. 641–664.

Demyanov, V. F. [2002]. The rise of nonsmooth analysis: Its main tools. *Cybernetics and Systems Analysis*, 38(4), pp. 527–547.

Demyanov, V. F., Bagirov, A. M., and Rubinov, A. M. [2002]. A method of truncated codifferential with application to some problems of cluster analysis. *Journal of Global Optimization*, 23(1), pp. 63–80.

Dreyer, D. R. and Overton, M. L. [1998]. Two heuristics for the Euclidean Steiner tree problem. *Journal of Global Optimization*, 13(1), pp. 95–106.

Drezner, Z. and Wesolowsky, G. O. [1991]. The Weber problem on the plane with some negative weights. *INFOR*, 29(2), pp. 87–99.

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

Eckhardt, U. [1980]. Weber's problem and Weiszfeld's algorithm in general spaces. *Mathematical Programming*, 18(1), pp. 186–196.

Ellaia, R. and Hiriart-Urruty, J.-B. [1986]. The conjugate of the difference of convex functions. *Journal of Optimization Theory and Applications*, 49(3), pp. 493–498.

Faraut, J. and Korányi, A. [1994]. *Analysis on Symmetric Cones*. Oxford University Press.

Faybusovich, L. [1997a]. Euclidean Jordan algebras and interior-point algorithms. *Positivity*, 1(4), pp. 331–357.

Faybusovich, L. [1997b]. Linear systems in Jordan algebras and primal-dual interior-point algorithms. *Journal of Computational and Applied Mathematics*, 86(1), pp. 149–175.

Fiacco, A. V. and McCormick, G. P. [1968]. *Nonlinear programming: Sequential unconstrained minimization*. Classics in Applied Mathematics. SIAM. (1990 republication).

Fletcher, R., Gould, N. I. M., Leyffer, S., Toint, P. L., and Wächter, A. [2002]. Global convergence of a trust-region SQP-filter algorithm for general nonlinear programming. *SIAM Journal on Optimization*, 13(3), pp. 635–659.

Fletcher, R. and Leyffer, S. [2002]. Nonlinear programming without a penalty function. *Mathematical Programming*, 91, pp. 239–269.

Forsgren, A., Gill, P. E., and Wright, M. H. [2002]. Interior methods for nonlinear optimization. *SIAM Review*, 44(4), pp. 525–597.

Gilbert, E. N. and Pollak, H. O. [1968]. Minimal Steiner trees. *SIAM Journal on Applied Mathematics*, 16(1), pp. 1–29.

Helsgaun, K. [2000]. An effective implementation of the Lin-Kernighan traveling salesman heuristic. *European Journal of Operational Research*, 126(1), pp. 106–130.

Hiriart-Urruty, J.-B. [1984]. Generalized differentiability, duality and optimization for problems dealing with differences of convex functions. In *Convexity and Duality in Optimization: Proceedings of the Symposium on Convexity and Duality in Optimization Held at the University of Groningen, the Netherlands, June 22, 1984*, number 256 in Lecture notes in Economics and Mathematical Systems, pp. 37–70. Springer.

Hiriart-Urruty, J.-B. [1986]. A general formula on the conjugate of the difference of functions. *Canadian Mathematical Bulleting*, 29(4), pp. 482–485.

Hiriart-Urruty, J.-B. [1988]. From convex optimization to non convex optimization, Part I: Necessary and sufficent conditions for global optimality. In *Nonsmooth Optimization and Related Topics*, edited by Clarke, F., Demyanov, V., and Giannessi, F., pp. 219–239. Plenum Press.

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

Hiriart-Urruty, J.-B. and Lemaréchal, C. [1993]. *Convex analysis and minimization algorithms I-II*. Springer.

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

Horst, R. and Pardolos, P. M. (editors) [1995]. *Handbook of Global Optimization*. Kluwer Academic Publishers.

Johnson, D. S. and McGeoch, L. A. [1997]. The traveling salesman problem: A case study in local optimization. In *Local Search in Combinatorial Optimization*, edited by Aarts, E. H. L. and Lenstra, J. K., pp. 215–310. John Wiley and Sons.

Johnson, D. S. and McGeoch, L. A. [2002]. Experimental analysis of heuristics for the STSP. In *The Traveling Salesman Problem and Its Variations*, edited by Gutin, G. and Punnen, A. P., pp. 369–443. Springer.

Jones, P. W. [1990]. Rectifiable sets and the traveling salesman problem. *Inventiones Mathematicae*, 102(1), pp. 1–15.

Karmarkar, N. [1984]. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4), pp. 373–395.

Kearfott, R. B. and Kreinovich, V. [2005]. Beyond convex? Global optimization is feasible only for convex objective functions: A theorem. *Journal of Global Optimization*, 33(4), pp. 617–624.

Koecher, M. [1999]. *The Minnesota notes on Jordan algebras and their applications*, volume 1710 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin.

Kojima, M., Mizuno, S., and Yoshise, A. [1991]. An $O(\sqrt{n}L)$ iteration potential reduction algorithm for linear complementarity problems. *Mathematical Programming*, 50(1), pp. 331–342.

Kuhn, H. W. [1973]. A note on Fermat's problem. *Mathematical Programming*, 4(1), pp. 98–107.

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

Kärkkäinen, T. and Äyrämö, S. [2005]. On computation of spatial median for robust data mining. In *Proceedings of EUROGEN 2005*, edited by Schilling, R., Haase, W., Periaux, J., and Baier, H. FLM, TU Munich, Munich, Germany.

Lerman, G. [2003]. Quantifying curvelike structures of measures by using L_2 Jones quantities. *Communications on Pure and Applied Mathematics*, 56, pp. 1294–1365.

Litke, J. D. [1984]. An improved solution to the traveling salesman problem with thousands of nodes. *Communications of the ACM*, 27(12), pp. 1227–1236.

Martínez-Legaz, J.-E. and Seeger, A. [1992]. A formula on the approximate subdifferential of the difference of convex functions. *Bulletin of the Australian Mathematical Society*, 45(1), pp. 37–41.

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

Mitrinović, D. S. [1970]. *Analytic Inequalities*. Springer-Verlag.

Monteiro, R. D. C. and Tsuchiya, T. [2000]. Polynomial convergence of primal-dual algorithms for the second-order cone program based on the MZ-family of directions. *Mathematical Programming*, 88(1), pp. 61–83.

Morris, J. G. [1981]. Convergence of the Weiszfeld algorithm for Weber problems using a generalized "distance" function. *Operations Research*, 29(1), pp. 37–48.

Muramatsu, M. [2002]. On a commutative class of search directions for linear programming over symmetric cones. *Journal of Optimization Theory and Applications*, 112(3), pp. 595–625.

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

Nesterov, Y. and Nemirovskii, A. [1994]. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM Studies in Applied Mathematics. SIAM.

Nesterov, Y. E. and Todd, M. J. [1997]. Self-scaled barriers and interior-point methods for convex programming. *Mathematics of Operations Research*, 22(1), pp. 1–42.

Ostresh, L. M., Jr. [1978]. On the convergence of a class of iterative methods for solving the Weber location problem. *Operations Research*, 26(4), pp. 597–609.

Pataki, G. [1996]. Cone-LPs and semidefinite programs: Geometry and a simplex-type method. In *Integer Programming and Combinatorial Optimization*, volume 1084 of *Lecture Notes in Computer Science*, pp. 162–174. Springer.

Penot, J.-P. [1978]. Calcul sous-différentiel et optimisation. *Journal of Functional Analysis*, 27(2), pp. 248–276.

Penot, J.-P. [1998]. On the minimization of difference functions. *Journal of Global Optimization*, 12, pp. 373–382.

Peyton Jones, S. et al. [2003]. The Haskell 98 language and libraries: The revised report. *Journal of Functional Programming*, 13(1), pp. 0–255.

Polak, P. and Wolansky, G. [2007]. The lazy travelling salesman problem in $\mathbb{R}^2$. *ESAIM: Control, Optimization and Calculus of Variations*, 13(3), pp. 538–552.

Potra, F. A. and Wright, S. J. [2000]. Interior-point methods. *Journal of Computational and Applied Mathematics*, 124(1-2), pp. 281–302.

Puerto, J. and Rodríguez-Chía, A. M. [1999]. Location of a moving service facility. *Mathematical Methods of Operations Research*, 49(3), pp. 373–393.

Puerto, J. and Rodríguez-Chía, A. M. [2006]. New models for locating a moving service facility. *Mathematical Methods of Operations Research*, 63(1), pp. 31–51.

Qi, L., Sun, D., and Zhou, G. [2002]. A primal-dual algorithm for minimizing a sum of Euclidean norms. *Journal of Computational and Applied Mathematics*, 138, pp. 127–250.

Reinelt, G. [1991]. TSPLIB—A traveling salesman problem library. *ORSA Journal on Computing*, 3(4), pp. 376–384.

Rockafellar, R. T. [1966]. Level sets and continuity of conjugate convex functions. *Transactions of the American Mathematical Society*, 123(1), pp. 46–63.

Rockafellar, R. T. [1972]. *Convex Analysis*. Princeton University Press.

Rockafellar, R. T. and Wets, R. J.-B. [1998]. *Variational Analysis*. Springer.

Schmieta, S. H. and Alizadeh, F. [2001]. Associative and Jordan algebras, and polynomial time interior-point algorithms for symmetric cones. *Mathematics of Operations Research*, 26(3), pp. 543–564.

Schmieta, S. H. and Alizadeh, F. [2003]. Extension of primal-dual interior point algorithms to symmetric cones. *Mathematical Programming*, 96(3), pp. 409–438.

Selim, S. Z. and Ismail, M. A. [1984]. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(1).

Teboulle, M. [2007]. A unified continuous optimization framework for center-based clustering methods. *Journal of Machine Learning Research*, 8, pp. 65–102.

Todd, M. J. and Ye, Y. [1990]. A centered projective algorithm for linear programming. *Mathematics of Operations Research*, 15(3), pp. 508–529.

Tuy, H. [1995]. D.C. optimization: Theory, methods and algorithms. In *Handbook of Global Optimization*, edited by Horst, R. and Pardolos, P. M., pp. 149–216. Kluwer Academic Publishers.

Ulbrich, M., Ulbrich, S., and Vicente, L. N. [2004]. A globally convergent primal-dual interior point filter method for nonconvex nonlinear programming. *Mathematical Programming*, 100, pp. 379–410.

Üster, H. and Love, R. F. [2000]. The convergence of the Weiszfeld algorithm. *Computers and Mathematics with Applications*, 40, pp. 443–451.

Valkonen, T. [2006]. Convergence of a SOR-Weiszfeld type algorithm for incomplete data sets. *Numerical Functional Analysis and Optimization*, 27(7–8), pp. 931–952. doi:10.1080/01630560600791213.

Valkonen, T. [2008a]. An errata to: Convergence of a SOR-Weiszfeld type algorithm for incomplete data sets. *Numerical Functional Analysis and Optimization*, 29(9–10), pp. 1201–1203. doi:10.1080/01630560802292028.

Valkonen, T. [2008b]. Optimality and sensitivity of DC functions. Submitted to Journal of Global Optimization.

Valkonen, T. [2008c]. A primal-dual interior point method for diff-convex problems on symmetric cones. Submitted to Mathematics of Operations Research.

Valkonen, T. and Kärkkäinen, T. [2008a]. Clustering and the perturbed spatial median. Submitted to Computer and Mathematical Modelling.

Valkonen, T. and Kärkkäinen, T. [2008b]. Continuous reformulations and heuristics for the Euclidean travelling salesperson problem. *ESAIM: Control, Optimization and Calculus of Variations*. doi:10.1051/cocv:2008056. URL `http://www.esaim-cocv.org/`. Published online (E-first).

Wächter, A. and Biegler, L. T. [2005]. Line search filter methods for nonlinear programming: Motivation and global convergence. *SIAM Journal on Computation*, 16(1), pp. 1–31.

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

Xue, G. and Ye, X. [1997]. An efficient algorithm for minimizing a sum of Euclidean norms with applications. *SIAM Journal on Optimization*, 7(4), pp. 1017–1036.

Yamashita, H. and Yabe, H. [2005]. A primal-dual interior point method for nonlinear optimization over second-order cones. Technical report, Mathematical Systems, Inc.

Äyrämö, S. [2006]. *Knowledge Mining Using Robust Clustering*. Number 63 in Jyväskylä Studies in Computing. University of Jyväskylä. Ph.D Thesis.

YHTEENVETO (FINNISH SUMMARY)

Tässä työssä tutkitaan optimointiongelmia, joissa kohdefunktio voidaan esittää Euklidisten etäisyyksien niin kutsuttuna diff-konveksina yhdistelmänä, eli konveksien funktioiden erotuksena. Työn tulokset jakautuvat neljään aihealueeseen: yleinen diff-konveksien funktioiden teoria, Weiszfeldin optimointimenetelmän laajennokset, sisäpistemenetelmät, sekä sovellukset sijaintiongelmiin. Näissä sovelluksissa tavoitteena on yhden tai useamman pisteen sijoittaminen (Euklidisessa) avaruudessa optimaalisesti määritellyn etäisyyksistä riippuvan ehdon mukaan.

Yleisen teorian alueella työssä esitetään uusia tuloksia optimaalisuusehtoihin liittyen, sekä näihin tuloksiin läheisesti liittyvää herkkyysanalyysiä. Lisäksi työssä tutkitaan funktioiden tasojoukkojen rajoittuneisuutta, sekä erään symmetrisiin kartioihin liittyvän diff-konveksien funktioiden luokan sisäistä rakennetta.

Näihin rakenneanalyyseihin pohjautuen työssä laajennetaan tähän funktioluokkaan sisäpistemenetelmiä lineaarisesta optimoinnista symmetrisillä kartiorajoitteilla. Työssä todistetaan paikallinen konvergenssi, ja tutkitaan suodatinmenetelmiin pohjautuvaa globalisointistrategiaa.

Weiszfeldin menetelmä laajennetaan työssä niin kutsuttuun "spatiaalimediaaniin epätäydellisellä datalla ja häiriöillä". Tässä spatiaalimediaanin kohdefunktiosta on vähennetty konveksi häiriötermi, ja lisäksi käytetyt etäisyydet mallintavat datan vaillinaisuutta. Työssä tutkitaan menetelmän konvergenssia sekä sovelluksia sijaintiongelmiin.

Työssä lähinnä tarkasteltavat sijaintiongelmat liittyvät klusterointiin ja Euklidiseen kaupparatsun ongelmaan. Klusterointiongelmista tutkitaan perinteistä niin kutsuttua usean lähteen Weberin ongelmaa eli K-spatiaalimediaania. Lisäksi esitellään uusi klusterointiongelman monitavoitetulkintaan pohjautuva kohdefunktio klusteroinnille, ja tutkitaan esitettyjen menetelmien soveltamista tähän. Tämän jälkeen työssä osoitetaan, että Euklidisen kaupparatsun ongelman ratkaisu voidaan määritellä ratkaisuna kummankin edellä mainitun klusterointi kohdefunktiolla lisättynä sakolla polunpituudelle.

Työn pääpaino on teoreettinen, käytännöllis-numeerisen puolen jäädessä vähemmälle.