Discontinuous piecewise differentiable optimization I: theory

A.R. Conn, Marcel Mongeau

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

1992

25ème anniversaire

N° 1694

Programme 5
Traitement du Signal,
Automatique et Productique

DISCONTINUOUS PIECEWISE DIFFERENTIABLE OPTIMIZATION I : THEORY

Andrew Roger CONN
Marcel MONGEAU

Mai 1992

Discontinuous Piecewise Differentiable Optimization I: Theory

Andrew R. Conn
T. J. Watson Research Center
P. O. Box 218, Yorktown Heights, N. Y. 10598

Marcel Mongeau
Centre de recherches mathématiques
Université de Montréal, C. P. 6128, Succ. A, Montréal, Canada H3C 3J7

Abstract

A theoretical framework and a practical algorithm are presented to solve discontinuous piecewise linear optimization problems. A penalty approach allows one to consider such problems subject to a wide range of constraints involving piecewise linear functions. Although the theory is expounded in detail in the special case of discontinuous piecewise *linear* functions, it is straightforwardly extendable, using standard nonlinear programming techniques, to the *nonlinear* (discontinuous piecewise differentiable) situation to yield a first order algorithm.

This work is presented in two parts. We introduce the theory in this first paper. The descent algorithm which is elaborated uses active set and projected gradient approaches. It is a generalization of the ideas used by Conn to deal with nonsmoothness in the l_1 exact penalty function, and it is based on the notion of *decomposition* of a function into a smooth and a nonsmooth part.

In an accompanying paper, we shall tackle constraints via a penalty approach, we shall discuss the degenerate situation, the implementation of the algorithm, and numerical results will be presented.

1 Introduction

We consider the problem:

$$\begin{array}{rrcl} \inf & \tilde{f}(x) & & \\ \text{subject to} & f_i(x) & = & 0,\ i \in E \\ & f_i(x) & \geq & 0,\ i \in I, \end{array} \qquad (1)$$

where the index sets E and I are finite and disjoint and $\tilde{f}$ and f_i, $i \in E \cup I$ are a collection of (possibly discontinuous) *piecewise linear* functions that map $I\!R^n$ to $I\!R$. A piecewise linear

Optimisation discontinue différentiable par morceaux I: Théorie

Andrew Roger Conn
T. J. Watson Research Center
P. O. Box 218, Yorktown Heights, N. Y. 10598

Marcel Mongeau
Centre de recherches mathématiques
Université de Montréal, C. P. 6128, Succ. A, Montréal, Canada H3C 3J7

Résumé

Un cadre théorique et un algorithme pratique sont présentés pour résoudre les problèmes d'optimisation linéaire par morceaux. L'utilisation d'une fonction de pénalité permet de considérer de tels problèmes sujets à des contraintes impliquant des fonctions linéaires par morceaux. Quoique la théorie soit développée pour le cas particulier des fonctions *linéaires* par morceaux, elle est facilement généralisable, en utilisant des techniques usuelles de programmation non-linéaire, au cas *non-linéaire* (différentiable par morceaux) conduisant à un algorithme de premier ordre.

Ce travail est présenté en deux parties, la théorie est présentée dans le présent rapport de recherche. L'algorithme de descente élaboré utilise les approches de contraintes actives et de gradient projeté. Notre approche généralise des idées utilisées par Conn pour traiter les fonctions non-lisses propres aux méthodes de fonction de pénalité exacte l_1 et est basée sur la notion de *décomposition* d'une fonction en une partie lisse et une partie non-lisse.

Dans un second article, nous traiterons les contraintes à l'aide d'une méthode de pénalité, nous discuterons du cas dégénéré, de l'implémentation de l'algorithme et des résultats numériques seront présentés.

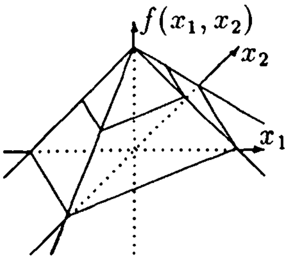

Figure 1: Example of a piecewise linear function f

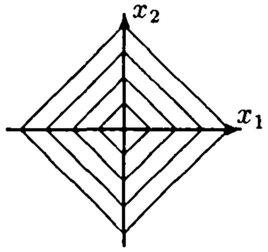

Figure 2: Level curves of f

function f is a function whose derivative is defined everywhere except over a subset of a finite number of hyperplanes and which is linear on the complement of the finite set of hyperplanes. A *ridge* is a specified hyperplane

$$\{x \in \mathbb{R}^n : a^T x = b\}, \ a \in \mathbb{R}^n, \ b \in \mathbb{R},$$

containing points where the derivative of f is not defined. We are concerned with finding a *local* infimum of the above optimization problem.

A simple example of such a piecewise linear function is given by the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined as follows:

$$f(x_1, x_2) = \begin{cases} -x_1 - x_2 & \text{if } x_1 \geq 0 \text{ and } x_2 \geq 0, \\ x_1 - x_2 & \text{if } x_1 < 0 \text{ and } x_2 \geq 0, \\ -x_1 + x_2 & \text{if } x_1 \geq 0 \text{ and } x_2 < 0, \\ x_1 + x_2 & \text{otherwise.} \end{cases} \tag{2}$$

The graph of f is a square base pyramid as shown on figure 1 while figure 2 shows level curves of f. We can consider the two lines $x_1 = 0$ and $x_2 = 0$ as being the set of ridges of f.

In [39], Zowe notes that, since they are not descent methods, subgradient methods (developed before the mid 70's) yield very poor convergence rates. He then motivates the *bundle* concept (see [29]) as an alternative direct approach to nondifferentiable optimization, overcoming some of the drawbacks of the subgradient methods. Bundle methods however involve a more complicated structure which demands a more sophisticated implementation. Other methods, such as [3, 12, 23, 24, 32, 33, 37], deal with specially structured nonsmooth functions, for example the minimax problem, the l_1 problem and elementary compositions of such problems. They are able to exploit the structure explicitly and consequently these methods are significantly more efficient than bundle methods (having 2-step superlinear rates of convergence [6, 12], for instance). Nevertheless, even for these problems, if they are extremely large, bundle methods, which can handle very general problems, may be more appropriate.

The idea of decomposition introduced in this paper will permit a broader generalization of Conn's approach.

Fourer [19, 20, 21] derived an extension of the simplex method for linear programming to (continuous) *monotropic* (i.e. convex, separable and linearly constrained) piecewise linear programming. His computationally practical algorithm is more efficient than indirect approaches that rely on transformation of piecewise linear programs to equivalent linear programs. See also [15, 16] which describe an extension to permit convex separable terms in the constraints.

Considering the optimization of functions which are nonsmooth and even discontinuous is motivated by applications in VLSI and floorplanning problems [4, 36], plant layout [31], batch production [26], switching regression [35], discharge allocation for hydro-electric generating stations [27], fixed-charge problems [17, 25, 34], for example.

Leaving aside the heuristic methods on which many people facing practical discontinuous optimization problems rely, previous work on discontinuous optimization includes smoothing algorithms. The smoothing algorithms express discontinuities by means of a step function, and then they approximate the step function by a function which is not only continuous but moreover smooth so that the resulting problem can be solved by a gradient technique. Both Imo & Leech [26] and Zang [38] developed methods in which the objective function is replaced only in the neighbourhood of the discontinuities. A drawback of these methods, not to mention the potential numerical instability when we want this neighbourhood to be small, is the cost of evaluating the smoothed functions.

One of the most important applications of nonsmooth optimization is in nonlinear programming. It arises when solving a smooth constrained optimization problem via minimizing an exact penalty function. We present an algorithm for discontinuous optimization based on ideas used in algorithms solving this very particular instance of (continuous) nonsmooth optimization problems. More specifically, the theory developed in this thesis is inspired by Conn's approach [8] to this problem. The discontinuous piecewise linear algorithm to be introduced here is thus a generalization of his algorithm.

In this paper we develop a theory to tackle the discontinuous piecewise linear optimization problem (1). We start from the idea used by Conn [8] that was applied to the nonsmooth problem resulting from the reduction of a constrained optimization problem to an unconstrained one via an l_1 exact penalty function. In the second part of this paper [13], we shall also use this same penalty approach in order to solve problem (1), that is we consider the unconstrained function

$$f_\gamma(x) \equiv \gamma \tilde{f}(x) + \sum_{i\in E} |f_i(x)| - \sum_{i\in I} \min[0, f_i(x)] \tag{3}$$

for a succession of decreasing positive values of the *penalty parameter* γ. The function f_γ is clearly a piecewise linear function, whenever the functions $\tilde{f}$ and f_i's are themselves piecewise linear. This penalty approach allows us to concentrate on the *unconstrained* nonsmooth optimization problem: $\inf_x f(x)$, where f is a (possibly discontinuous) piecewise linear function.

The next section sets the terminology required for the reading of the paper and presents the concepts of *activities* and *restricted gradient*, which are fundamental to the theory to be presented.

We introduce in section 3 the definition of *decomposition* of a continuous piecewise linear function into a smooth function and a sum of single-ridged functions, the theorem stating that such a decomposition can always be found at a non-degenerate point, the optimality conditions and the algorithm.

The following section extends the theory to the *discontinuous* piecewise linear situation by generalizing the definition of decomposition, then the decomposition theorem and hence the optimality conditions and the algorithm.

The penultimate section explains how our work can be straightforwardly extended to yield a first-order algorithm for the nonlinear case—the general (possibly discontinuous) *piecewise differentiable* situation.

Section 6 concludes this first part of our paper.

Details of implementation, the constrained case, degeneracy, a discussion on how to deal with singular points called *contact points*, and numerical results are deferred to the second part [13] of this paper.

2 Terminology

We define the function *sign* as follows:

$$\mathrm{sign}(x) \equiv \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{if } x < 0, \\ 0 & \text{otherwise.} \end{cases}$$

If V is a subspace in $\mathbb{R}^n$, then

$$V^\perp \equiv \{x \in \mathbb{R}^n : v^T x = 0, \text{ for all } v \in V\}$$

is termed the *orthogonal complement* of V. The *range space* of an $m \times n$ matrix M is the set of vectors that can be written as a linear combination of the columns of M. The null space of M, denoted by $\mathcal{N}(M)$, is defined by

$$\mathcal{N}(M) \equiv \{x \in \mathbb{R}^n : Mx = \vec{0}\}$$

(it is the orthogonal complement of the range space of M^T).

Let f map $\mathbb{R}^n$ to $\mathbb{R}$. The usual (one-sided) *directional derivative* or the *first order change* of f at $x \in \mathbb{R}^n$ in the direction $v \in \mathbb{R}^n$ is

$$f'(x; v) \equiv \lim_{t \to 0^+} \frac{f(x + tv) - f(x)}{t}$$

when this limit exists.

An $n \times n$ matrix P is called an *orthogonal projector* if P is *symmetric* (i.e. $P^T = P$) and *idempotent* ($P \cdot P = P$). For P to be an orthogonal projector onto a subspace $V \subseteq \mathbb{R}^n$

means that P projects a vector $v \in V$ onto itself ($v \in V$ implies $Pv = v$) and P projects a vector of the orthogonal complement of V onto $\vec{0}$ ($v \in V^{\perp}$ implies $Pv = \vec{0}$). If A is an $n \times m$ matrix with linearly independent columns, one can show that the orthogonal projector onto $\mathcal{N}(A^T)$ is given by

$$P \equiv I - A(A^T A)^{-1} A^T,$$

where I is the $n \times n$ identity matrix.

If $\{x \in I\!\!R^n : a^T x - b = 0\}$, $a \in I\!\!R^n$, $b \in I\!\!R$ is a ridge of f, then, by abuse of language, we shall call $a^T x - b$ a ridge of f. We identify two ridges $a_1^T x - b_1$ and $a_2^T x - b_2$ if $\{x \in I\!\!R^n : a_1^T x - b_1 = 0\} = \{x \in I\!\!R^n : a_2^T x - b_2 = 0\}$. Let $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$ be the ridges of f, where $\mathcal{R}$ is a finite index set. We say that a ridge $a^T x - b$ is *active* at $\hat{x}$ if $a^T \hat{x} - b = 0$. Let $\mathcal{A}(\hat{x}) \subseteq \mathcal{R}$ be the (finite) index set of the ridges that are active at the current point $\hat{x}$, and $A(\hat{x})$ be the matrix having as columns the gradients of the ridges which are active at $\hat{x}$. Hence, $\{a_i\}_{i \in \mathcal{A}(\hat{x})}$ denotes the column vectors of the *matrix of activities* $A(\hat{x})$. We shall use A and $\mathcal{A}$ rather than $A(\hat{x})$ and $\mathcal{A}(\hat{x})$, when it is clear from the context which is the current point. Also, when there is no confusion possible, we shall often talk about activity k or ridge k, meaning the ridge indexed by $k \in \mathcal{R}$.

A direction $d \in \mathcal{N}(A^T(\hat{x}))$ is said to *preserve* each activity $i \in \mathcal{A}(\hat{x})$, since for each $i \in \mathcal{A}(\hat{x})$ we have, $a_i^T(\hat{x} + \alpha d) - b_i = a_i^T \hat{x} - b_i = 0$. If $\mathcal{A}(\hat{x}) \neq \emptyset$, then $\nabla f(\hat{x})$ is not necessarily defined. The problem comes from the fact that we cannot talk about the gradient of the function at $\hat{x}$ since there is no vector $g \in I\!\!R^n$ such that $g^T d$ is the first order change of f in direction d, for any $d \in I\!\!R^n$. Thus, we cannot use, as in the smooth situation, the negative gradient direction as a descent direction. We shall first consider directions $d \in \mathcal{N}(A^T)$ such that d is a descent direction. We term any $n \times 1$ vector $g_{\hat{x}}$ such that

$$f'(\hat{x}; d) = g_{\hat{x}}^T d, \quad \text{for all } d \in \mathcal{N}(A^T)$$

a *restricted gradient* of f at $\hat{x}$, because it is the gradient of the restriction of f to the space $\mathcal{N}(A^T(\hat{x}))$.

To illustrate these concepts, consider the following simple example:

$$\min_{x \in I\!\!R^2} \; -f(x),$$

$$\text{subject to } \; x_1 + x_2 \geq 0,$$

where $f(x)$ is defined by (2). It yields the unconstrained optimization problem: $\min_{x \in I\!\!R^2} f_\gamma(x)$, where

$$f_\gamma(x) = -\gamma f(x) - \min(0, x_1 + x_2) \tag{4}$$

and $\gamma > 0$. The derivative of the penalty function f_γ is not defined only over the three ridges

$$a_i^T x - b_i, \; i \in \mathcal{R} \equiv \{1, 2, 3\},$$

where

$$a_1 = (1,0)^T, \;\; a_2 = (0,1)^T, \;\; a_3 = (1,1)^T, \tag{5}$$

and $b_i = 0,\ i \in \mathcal{R}$. At the point $\tilde{x}^T = (0,0)$, the three ridges are active, i.e. $\mathcal{A}(\tilde{x}) = \{1,2,3\}$ and hence the matrix of activities is

$$A(\tilde{x}) = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

At $\hat{x}^T = (0,1)$, we have $\mathcal{A}(\hat{x}) = \{1\}$ and

$$A(\hat{x}) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

It will be useful for what follows to introduce the functions $\{\sigma_i\}_{i \in \mathcal{R}}$ defined by:

$$\sigma_i(x) \equiv \mathsf{sign}(a_i^T x - b_i) \quad , i \in \mathcal{R}.$$

We denote $\sigma(x)$ the $|\mathcal{R}| \times 1$ vector whose ith component is $\sigma_i(x)$, and $3^{|\mathcal{R}|}$ is the set of all possible such vectors. Given a (possibly discontinuous) piecewise linear function f defined over $I\!R^n$ and the set $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$ of its ridges, a *cell* of f is a non-empty set $C \subseteq I\!R^n$ such that for all $x, y \in C$ we have $\sigma_i(x) = \sigma_i(y) \neq 0$, for all $i \in \mathcal{R}$. Thus, f is differentiable over a cell.

For example, the interior of the four orthants constitute a set of cells of function f given by (2).

Let $\sigma \in 3^{|\mathcal{R}|}$, $i \in \mathcal{R}$ and σ_i be the ith component of vector σ. A set

$$\{x : \sigma_i(x) = 0 \text{ and } \sigma_j(x) = \sigma_j,\ j \in \mathcal{R} \setminus \{i\}\}$$

is called a *segment* of ridge i.

The half-line $\{x \in I\!R^2 : x_2 = 0 \text{ and } x_1 > 0\}$ is a segment of the ridge $x_2 = 0$ (for the function f given by (2)).

Figure 3 shows the values of the gradient of the penalty function, f_γ, given by (4) when restricted to each cell. One can easily verify that a restricted gradient of f_γ at $\hat{x}^T$ is $g_{\hat{x}} = (\gamma, \gamma)$. Indeed, along any direction $d \in I\!R^2$ preserving activity 1 (i.e. such that $a_1^T d = 0$) we have $f'(\hat{x}; d) = g_{\hat{x}}^T d$.

Given a point $\hat{x} \in I\!R^n$ and a line $L = \{x \in I\!R^n : x = \hat{x} + \alpha d, \alpha \in I\!R\}$, we say that x_0 is a *breakpoint* of f along d when at least one ridge i of f is active at x_0 with $a_i^T d \neq 0$ (i.e. i is not active at $\hat{x}$). In our example, starting from $\hat{x}^T = (0,1)$, $x_0^T = (-1,1)$ is a breakpoint of f_γ along direction $d = (-1,0)^T$, as we "hit" ridge 3 at x_0 (see figure 3).

3 Continuous Piecewise Linear Optimization

We will assume that a continuous piecewise linear function, f, is given under the following form: The ridges, $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$, where $\mathcal{R}$ is a finite index set, of f are given, i.e. we are given: $a_i \in I\!R^n$ and $b_i \in I\!R$, for each $i \in \mathcal{R}$. At any point $\hat{x} \in I\!R^n$, the set of ridges which are active at $\hat{x}$, $\mathcal{A}(\hat{x})$, is given. Finally, consider the set

$$C_{\hat{x}}^P \equiv \{x \in B(\hat{x}) : a_i^T x > b_i, i \in P \text{ and } a_i^T x < b_i, i \in \mathcal{A}(\hat{x}) \setminus P\},$$

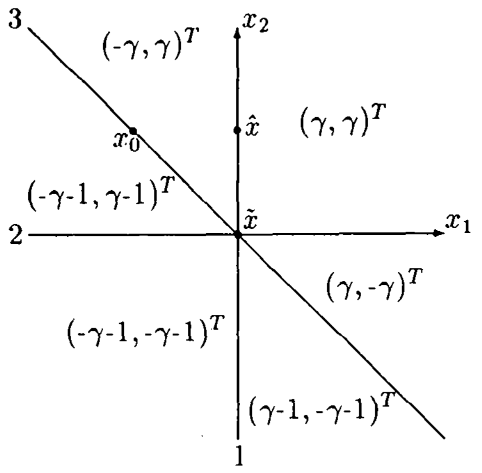
3
$(-\gamma,\gamma)^T$
x_2
$\hat{x}$
$(\gamma,\gamma)^T$
x_0
$(-\gamma\text{-}1,\gamma\text{-}1)^T$
$\tilde{x}$
2
x_1
$(\gamma,-\gamma)^T$
$(-\gamma\text{-}1,-\gamma\text{-}1)^T$
$(\gamma\text{-}1,-\gamma\text{-}1)^T$
1

Figure 3: Values of the gradient of f_γ over the six cells

where $B(\hat{x})$ is a neighbourhood of $\hat{x}$ and $P \subseteq \mathcal{A}(\hat{x})$. If $B(\hat{x})$ is small enough, f is linear over $C_{\hat{x}}^P$, i.e. $f(x) = f(\hat{x}) + x^T c_{\hat{x}}^P$, $x \in C_{\hat{x}}^P$, for some vector $c_{\hat{x}}^P \in I\!R^n$. We shall consider that at any point $\hat{x} \in I\!R^n$ and for any $P \subseteq \mathcal{A}(\hat{x})$ such that $C_{\hat{x}}^P \neq \emptyset$, we can obtain the vector $c_{\hat{x}}^P$. (Hence, we are assuming that more information on the structure of the objective function is available than, for example, in a bundle method, which assumes that only one element of the subdifferential is known at any point.)

3.1 Decomposition

Definition 1 *Let $f : I\!R^n \longrightarrow I\!R$ be a continuous piecewise linear function with ridges $\{a_i^T x - b_i\}_{i\in\mathcal{R}}$, where $\mathcal{R}$ is a finite index set. Let $\hat{x} \in I\!R^n$, $g_{\hat{x}} \in I\!R^n$ and $\psi_{\hat{x}}$ be a function defined on $I\!R^n$ such that we have:*

$$f(x) = f(\hat{x}) + g_{\hat{x}}^T(x - \hat{x}) + \psi_{\hat{x}}(x),$$

for all x in some neighbourhood, $B(\hat{x})$, of $\hat{x}$, where

$$\psi_{\hat{x}}(x) = \sum_{i\in\mathcal{A}(\hat{x})} \nu_{\hat{x}}^i \min(0, a_i^T(x - \hat{x}))$$

for some scalars $\{\nu_{\hat{x}}^i\}_{i\in\mathcal{A}(\hat{x})}$.

We say that $g_{\hat{x}}, \{\nu_{\hat{x}}^i\}_{i\in\mathcal{A}(\hat{x})}$ is a decomposition of f (into a smooth function and a sum of continuous functions having a single ridge) at $\hat{x}$.

To illustrate, one way of decomposing function f_γ given by (4) (see figure 3) at $\tilde{x} = (0,0)^T$ could be the following:

$$f_\gamma(x) = g_{\tilde{x}}^T x + \sum_{i\in\mathcal{A}(\tilde{x})} \nu_{\tilde{x}}^i \min(0, a_i^T x), \tag{6}$$

for all $x \in I\!R^2$, where $\mathcal{A}(\tilde{x}) = \{1,2,3\}$; a_i, $i = 1,2,3$ are defined by (5); and

$$\begin{aligned} g_{\hat{x}}^T &= (\gamma,\gamma); \\ \nu_{\hat{x}}^1 &= -2\gamma; \\ \nu_{\hat{x}}^2 &= -2\gamma; \\ \nu_{\hat{x}}^3 &= -1. \end{aligned}$$

We leave out the subscript 'γ' in $g_{\hat{x}}$ and $\nu_{\hat{x}}^i$ for notational simplicity.

We shall shortly give a practical way to find a decomposition at a given point $\hat{x}$. In the above example, we can easily provide "simultaneously" a decomposition for each $\hat{x}$ in the domain of f_γ, that is, a decomposition such that $\nu_{\hat{x}}^i = \nu_{\tilde{x}}^i$ at any points $\hat{x}, \tilde{x}$, $i \in \mathcal{R}$. We shall later deal with instances, in degenerate situations, where the scalars $\{\nu_{\hat{x}}^i\}_{i \in \mathcal{R}}$ do depend upon $\hat{x}$. In fact, even in these instances, they depend only on $\sigma(\hat{x})$. Indeed, we shall see that we can define a decomposition with common $\{\nu_{\hat{x}}^i\}_{i \in \mathcal{R}}$ for all $\hat{x}$ which are on the same segment of a ridge.

When there exists a decomposition of a function f at a point $\hat{x} \in I\!R^n$, we say that f is *decomposable* at $\hat{x}$. Suppose that f is decomposable at a point $\hat{x}$, and let $g_{\hat{x}}, \{\nu_{\hat{x}}^i\}_{i \in \mathcal{A}(\hat{x})}$ be the decomposition. Choosing a direction that would not change the value of *the nonsmooth part* of f, $\psi_{\hat{x}}$, (for example, a direction preserving the current activities) and reduce that of *the smooth part*, $f(\hat{x}) + g_{\hat{x}}^T(x - \hat{x})$, (whenever possible) would be enough to obtain descent in f. We have, for $\alpha > 0$ small enough:

$$f(\hat{x} + \alpha d) = f(\hat{x}) + \alpha g_{\hat{x}}^T d + \alpha \sum_{i \in \mathcal{A}(\hat{x}) : a_i^T d < 0} \nu_{\hat{x}}^i a_i^T d.$$

Thus, restricting d to be in $\mathcal{N}(A^T(\hat{x}))$ we obtain

$$f(\hat{x} + \alpha d) = f(\hat{x}) + \alpha g_{\hat{x}}^T d.$$

Clearly, $g_{\hat{x}}$ is a restricted gradient of f at $\hat{x}$. Given a decomposition, $g_{\hat{x}}, \{\nu_{\hat{x}}^i\}_{i \in \mathcal{A}(\hat{x})}$, of f at $\hat{x}$, we shall call $g_{\hat{x}}$, by convention, *the* restricted gradient of f at $\hat{x}$. The next theorem will state that, given the set of ridges $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$ of f at $\hat{x}$, the decomposition of f at $\hat{x}$ and, hence, the restricted gradient of f at $\hat{x}$, are unique provided that the gradients of the ridges which are active at $\hat{x}$ are linearly independent.

Unless it is essential to specify the subscript '$\hat{x}$' in $g_{\hat{x}}, \{\nu_{\hat{x}}^i\}_{i \in \mathcal{A}(\hat{x})}$, we shall sometimes omit it by abuse of notation and talk only about a decomposition $g, \{\nu^i\}_{i \in \mathcal{A}}$ of f at $\hat{x}$.

Let A_{-k} be the matrix having as columns the gradients of the ridges which are active except for activity $k \in \mathcal{A}$. We call a direction $d \in \mathcal{N}(A_{-k}^T)$, $k \in \mathcal{A}$, such that $a_k^T d \neq 0$, a *single-dropping* direction or a direction *dropping* only activity k, since it preserves all activities but activity k (if $\mathcal{A} = \{k\}$, then $\mathcal{N}(A_{-k})$ is simply $I\!R^n$). We say that activity k is *dropped positively* if moreover $a_k^T d > 0$, and that activity k is *dropped negatively* when d is such that $a_k^T d < 0$. With a decomposition $g, \{\nu^i\}_{i \in \mathcal{A}}$ of f at $\hat{x}$, we know that there is a vector $v \in I\!R^n$ such that if d is a single-dropping direction, then the change in the nonsmooth part is $v^T d$, where v *is a multiple of the gradient of the ridge being dropped.* This observation will permit derivation of optimality conditions similar to Lagrange multiplier rules: a multiplier rule will be associated with each activity.

The minimization of some special cases of continuous piecewise linear functions have been studied before, for example, in l_1 data fitting or when using the l_1 exact penalty function to solve a linear programming problem [1, 8]. In the latter problem, we want to minimize a particular piecewise linear penalty function f_γ. The function f_γ is as given in (3), where $\tilde{f}$ and f_i, $i \in E \cup I$, are linear. Thus, the penalty function to minimize has the form

$$f_\gamma(x) = \gamma c^T x + \sum_{i \in E} |a_i^T x - b_i| - \sum_{i \in I} \min(0, a_i^T x - b_i), \tag{7}$$

for some $c \in I\!R^n$; $a_i \in I\!R^n$, $b_i \in I\!R$, $i \in E \cup I$. In order to solve this problem, Conn defined the restricted gradient of f_γ at x to be

$$g(x) = \gamma c + \sum_{i \in E} a_i \text{sign}(a_i^T x - b_i) - \sum_{i \in I} a_i s(a_i^T x - b_i),$$

where the function $s(x)$ is defined to take the value 1 if x is negative and 0 otherwise. This way, he could express f_γ, in a neighbourhood of a point $\hat{x}$, as the sum of a differentiable part, $f_\gamma(\hat{x}) + g(\hat{x})^T(x - \hat{x})$, and a nondifferentiable part,

$$\sum_{i \in E \cap \mathcal{A}(\hat{x})} |a_i^T x - b_i| - \sum_{i \in I \cap \mathcal{A}(\hat{x})} \min(0, a_i^T x - b_i).$$

Clearly the change of the nondifferentiable part of f_γ in a single-dropping direction d can easily be written as the inner product of d with a multiple of the gradient of the activity to be dropped. For example, if at the current point $\hat{x}$, $a_k^T x - b_k$ is active and d drops negatively this activity, then

$$f'_\gamma(\hat{x}, d) = (g(\hat{x}) + (-1)a_k)^T d.$$

If d drops k positively, then

$$f'_\gamma(\hat{x}, d) = \begin{cases} (g(\hat{x}) + 1a_k)^T d & \text{if } k \in E, \\ (g(\hat{x}) + 0a_k)^T d & \text{if } k \in I. \end{cases}$$

Given an arbitrary continuous piecewise linear function f, two questions remain: How to find a decomposition of f into a smooth function and a sum of continuous functions having a single ridge and, above all, does such a decomposition always exist? At first sight, the answer to the latter question seems unlikely to be yes. Indeed, if m ridges of f are active at $\hat{x}$, it means that there are 2^m cells in any small neighbourhood of $\hat{x}$, assuming $\{a_i\}_{i \in \mathcal{A}(\hat{x})}$ linearly independent. One can then wonder how could the vector $g_{\hat{x}}$ and the m scalars $\{\nu_{\hat{x}}^i\}_{i \in \mathcal{A}(\hat{x})}$, together with the gradients of the activities, $\{a_i\}_{i \in \mathcal{A}(\hat{x})}$, completely characterize the behaviour of f over the 2^m cells in the neighbourhood of $\hat{x}$. The next theorem gives a sufficient condition for a continuous piecewise linear function to be decomposable at a given point while the proposition following it gives a practical way to construct the decomposition.

Theorem 1 (Decomposition) *Let $f : I\!R^n \to I\!R$ be a continuous piecewise linear function with ridges $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$, where $\mathcal{R}$ is a finite index set, and let $\hat{x} \in I\!R^n$. If $\{a_i\}_{i \in \mathcal{A}(\hat{x})}$, the gradients of the ridges of f which are active at $\hat{x}$, are linearly independent, then f is decomposable at $\hat{x}$ and the decomposition is unique.*

To give an intuitive idea as to why the result is true, one should note that in constructing an arbitrary piecewise linear function having m linearly independent ridges active at $\hat{x}$, the requirement of the continuity of f reduces greatly the number of degrees of freedom. One cannot just assign arbitrary linear functions to each of the 2^m cells in the neighbourhood of $\hat{x}$. Consider for example the partition of $I\!R^2$ into four cells by the ridges $x_1 = 0$ and $x_2 = 0$. Defining a function over the second and fourth orthants determines completely a continuous piecewise linear function over the whole domain. In fact, assigning m linear functions to m different cells will be enough to determine the value of f on the $2^m - m$ remaining cells, if we require f to be continuous.

Proof: Let us suppose without loss of generality that $f(\vec{0}) = 0$, and $\hat{x} = \vec{0}$ (thus, $b_i = 0$ for all $i \in \mathcal{A}(\hat{x})$), and $\mathcal{A}(\hat{x}) \equiv \{1, 2, \dots, m\}$. In a neighbourhood, $B(\hat{x})$, of $\hat{x}$, function f has the form

$$f(x) = x^T c^J$$

whenever

$$\begin{array}{rrl} a_i^T x & \geq & 0, \;\; i \in J \\ \text{and} \;\; a_i^T x & < & 0, \;\; i \in \{1, 2, \dots, m\} \setminus J, \end{array}$$

for some vectors $c^J \in I\!R^n$ given for each $J \subseteq \{1, 2, \dots, m\}$. We can define an invertible linear transformation,

$$y = Lx,$$

such that

$$y_i \equiv a_i^T x, \;\; 1 < i \leq m$$

(and in the case where $m < n$,

$$y_i \equiv \tilde{a}_i^T x, \;\; m < i \leq n,$$

where the $\tilde{a}_i \in I\!R^n$ are such that $\{a_i, \; 1 \leq i \leq m\} \cup \{\tilde{a}_i, \; m < i \leq n\}$ are linearly independent). Thus, we can write f, in a neighbourhood, $B(\hat{x})$, of $\hat{x}$, using the form

$$f(x) = (L^{-1}y)^T c^J$$

whenever

$$\begin{array}{rrl} y_i & \geq & 0, \;\; i \in J \\ \text{and} \;\; y_i & < & 0, \;\; i \in \{1, 2, \dots, m\} \setminus J, \end{array}$$

where the $c^J \in I\!R^n$ are given for each $J \subseteq \{1, 2, \dots, m\}$.

Let

$$d^J \equiv L^{-T} c^J$$

for each $J \subseteq \{1, 2, \dots, m\}$. Then, in a neighbourhood of $\hat{x}$, $f(x) = h(Lx)$, where

$$h(y) = y^T d^J$$

whenever

$$\begin{aligned} y_i &\geq 0, \quad i \in J \\ \text{and} \quad y_i &< 0, \quad i \in \{1,2,\ldots,m\} \setminus J. \end{aligned}$$

Since the function f is continuous, h is continuous. The continuity of h yields, for $I, J \subseteq \{1,2,\ldots,m\}$:

$$d_i^I = d_i^J \text{ whenever } i \in I \cap J \text{ or } i \in \{1,2,\ldots,m\} \setminus (I \cup J) \text{ or } m < i \leq n. \tag{8}$$

Clearly, (8) is true if and only if for $I \subseteq \{1,2,\ldots,m\}$ and $1 \leq i \leq n$:

$$d_i^I = \begin{cases} d_i^{\{1,2,\ldots,m\}} & \text{if } i \in I, \\ d_i^{\emptyset} & \text{otherwise.} \end{cases}$$

Hence,

$$h(y) = \sum_{i=1}^{n} d_i y_i,$$

where

$$d_i = \begin{cases} d_i^{\{1,2,\ldots,m\}} & \text{if } y_i \geq 0, \\ d_i^{\emptyset} & \text{if } y_i < 0. \end{cases}$$

That is to say,

$$h(y) = y^T d^{\{1,2,\ldots,m\}} + \sum_{i=1}^{m} (d_i^{\emptyset} - d_i^{\{1,2,\ldots,m\}}) \min(0, y_i),$$

since for $m < i \leq n$, $d_i^{\emptyset} = d_i^{\{1,2,\ldots,m\}}$. Hence, in a neighbourhood of $\hat{x}$,

$$f(x) = x^T c^{\{1,2,\ldots,m\}} + \sum_{i=1}^{m} \nu^i \min(0, a_i^T x), \text{where } \nu^i \equiv d_i^{\emptyset} - d_i^{\{1,2,\ldots,m\}},$$

which is a decomposition of f at $\hat{x}$.

In order to prove the uniqueness of the decomposition, let $g, \{\nu^i\}_{i \in \mathcal{A}}$ and $g', \{\nu'^i\}_{i \in \mathcal{A}}$ be two decompositions of f at $\hat{x}$. By definition of the decomposition, we have for all x in some neighbourhood, $B(\hat{x})$, of $\hat{x}$:

$$f(x) = g^T x + \sum_{i \in \mathcal{A}} \nu^i \min(0, a_i^T x) = g'^T x + \sum_{i \in \mathcal{A}} \nu'^i \min(0, a_i^T x). \tag{9}$$

Using (9) with $x \in B(\hat{x})$ such that $a_i^T x \geq 0$ for all $i \in \mathcal{A}$ yields

$$g^T x = g'^T x$$

$$\text{or } (g - g')^T x = 0. \tag{10}$$

Equation (10) is true for any $x \in B(\hat{x})$ in the set

$$S \equiv \{x \in I\!R^n : a_i^T x \geq 0, \text{ for all } i \in \mathcal{A}\}.$$

In the case where $|\mathcal{A}| = n$, we show that $g - g' = \vec{0}$ by considering each of the (linearly independent) directions

$$d^k \equiv P_{-k}(a_k), \quad k \in \mathcal{A},$$

where P_{-k} is the orthogonal projector onto $\mathcal{N}(A_{-k}(\hat{x}))$. If $|\mathcal{A}| \equiv m < n$, then consider each of the directions

$$d^k \equiv \tilde{P}_{-k}(a_k), \quad k \in \mathcal{A} \text{ and}$$

$$d^k \equiv \tilde{P}_{-k}(\tilde{a}_k), \quad k \in \{1, 2, \ldots, n\} \setminus \mathcal{A},$$

where $\tilde{P}_{-k}$ is the orthogonal projector onto $\mathcal{N}(\tilde{A}_{-k}(\hat{x}))$, $\tilde{A}_{-k}(\hat{x})$ has as its columns the vectors $a_1, \ldots, a_m, \tilde{a}_{m+1}, \ldots, \tilde{a}_n$ and the $\tilde{a}_k$'s extend the a_k's to form a basis of the entire space.

We show that $\nu^i = \nu'^i$ for any $i \in \mathcal{A}(\hat{x})$ by using (9) with $x = -P_{-i}(a_i)$. □

Proposition 1 *Let $f : I\!R^n \to I\!R$ be a continuous piecewise linear function with ridges $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$, where $\mathcal{R}$ is a finite index set, and let $\hat{x} \in I\!R^n$. Assume that $\{a_i\}_{i \in \mathcal{A}(\hat{x})}$, the gradients of the ridges of f which are active at $\hat{x}$, are linearly independent and that, without loss of generality, f has the following form in a neighbourhood of $\hat{x}$:*

$$f(x) = f(\hat{x}) + (x - \hat{x})^T c^J \tag{11}$$

whenever

$$\begin{array}{rrcl} & a_i^T(x - \hat{x}) & \geq & 0, \quad i \in J \\ \textit{and} & a_i^T(x - \hat{x}) & < & 0, \quad i \in \mathcal{A}(\hat{x}) \setminus J, \end{array}$$

for some $c^J \in I\!R^n$ given for each $J \subseteq \mathcal{A}(\hat{x})$.

Then, $c^{\mathcal{A}(\hat{x})}$, $\{\lambda^i\}_{i \in \mathcal{A}(\hat{x})}$ is the decomposition of f at $\hat{x}$, where each of the scalars $\{\lambda^i\}_{i \in \mathcal{A}(\hat{x})}$ is such that

$$\lambda^i a_i = c^{J_i - i} - c^{J_i},$$

for any *given $J_i \subseteq \mathcal{A}(\hat{x})$ such that $i \in J_i$.*

Proof: Let $\{J_i\}_{i \in \mathcal{A}(\hat{x})}$ be such that $i \in J_i \subseteq \mathcal{A}(\hat{x})$, for each $i \in \mathcal{A}(\hat{x})$. Note first that since f is decomposable at $\hat{x}$, then there exists some decomposition $g, \{\nu^i\}_{i \in \mathcal{A}(\hat{x})}$ of f at $\hat{x}$, and hence, for any $J \subseteq \mathcal{A}(\hat{x})$ we have, by (11):

$$c^J = g + \sum_{k \in \mathcal{A}(\hat{x}) \setminus J} \nu^k a_k.$$

Thus, for any $J \subseteq \mathcal{A}(\hat{x})$ such that $i \in J$ we have

$$\begin{aligned} c^{J-i} - c^J &= g + \sum_{k \in \mathcal{A}(\hat{x}) \setminus (J-i)} \nu^k a_k - (g + \sum_{k \in \mathcal{A}(\hat{x}) \setminus J} \nu^k a_k) \\ &= \nu^i a_i. \end{aligned}$$

This is true in particular for $J = J_i$; that is to say

$$c^{J_i - i} - c^{J_i} = \nu^i a_i,$$

which implies that $\lambda^i a_i = \nu^i a_i$, and hence,

$$c^{J-i} - c^J = \lambda^i a_i, \tag{12}$$

for any $J \subseteq \mathcal{A}(\hat{x})$ such that $i \in J$.

Now let $\tilde{x}$ be an arbitrary point of a small neighbourhood of $\hat{x}$. In order to show that $c^{\mathcal{A}(\hat{x})}, \{\lambda^i\}_{i \in \mathcal{A}(\hat{x})}$ is a decomposition of f at $\hat{x}$, we would like to show that

$$f(\tilde{x}) = f(\hat{x}) + (\tilde{x} - \hat{x})^T c^{\mathcal{A}(\hat{x})} + \sum_{i \in \mathcal{A}(\hat{x})} \lambda^i \min(0, a_i^T(\tilde{x} - \hat{x})).$$

Let

$$K \equiv \{i \in \mathcal{A}(\hat{x}) : a_i^T(\tilde{x} - \hat{x}) \geq 0\}$$

and

$$K^c \equiv \{i \in \mathcal{A}(\hat{x}) : a_i^T(\tilde{x} - \hat{x}) < 0\}.$$

Suppose without loss of generality that

$$K^c \equiv \{1, 2, \ldots, l\},$$

and let

$$K_i \equiv K \cup \{1, 2, \ldots, i\},\ i \in K^c, \quad \text{and} \quad K_0 \equiv K.$$

We have

$$\begin{aligned}
f(\hat{x}) \quad & +(\tilde{x} - \hat{x})^T c^{\mathcal{A}(\hat{x})} + \sum_{i \in \mathcal{A}(\hat{x})} \lambda^i \min(0, a_i^T(\tilde{x} - \hat{x})) \\
= \quad & f(\hat{x}) + (\tilde{x} - \hat{x})^T c^{\mathcal{A}(\hat{x})} + \sum_{i=1}^{l} \lambda^i a_i^T(\tilde{x} - \hat{x}) \\
= \quad & f(\hat{x}) + (\tilde{x} - \hat{x})^T c^{\mathcal{A}(\hat{x})} + \sum_{i=1}^{l} (c^{K_i - i} - c^{K_i})^T(\tilde{x} - \hat{x}), \text{ using (12) with } J = K_i, \\
= \quad & f(\hat{x}) + (\tilde{x} - \hat{x})^T c^{\mathcal{A}(\hat{x})} + \sum_{i=1}^{l} (c^{K_{i-1}} - c^{K_i})^T(\tilde{x} - \hat{x}) \\
= \quad & f(\hat{x}) + (\tilde{x} - \hat{x})^T c^{\mathcal{A}(\hat{x})} + (c^{K_0} - c^{K_l})^T(\tilde{x} - \hat{x}) \\
= \quad & f(\hat{x}) + (\tilde{x} - \hat{x})^T c^{K}, \text{ since } K_l \equiv \mathcal{A}(\hat{x}) \text{ and } K_0 \equiv K, \\
= \quad & f(\tilde{x}),
\end{aligned}$$

by (11) and the definition of K. □

Thus, intuitively, we have for a continuous piecewise linear function f that, in the neighbourhood of a point $\hat{x}$ such that $\{a_i\}_{i \in \mathcal{A}(\hat{x})}$ are linearly independent, the difference between the gradient of f when f is restricted to a cell C_1, and the gradient of f when f is restricted to an "adjacent" cell C_2, i.e. C_1 and C_2 are "separated" only by *one* ridge, $a_i^T x - b_i$, is always a multiple of a_i, the gradient of the ridge. The factor involved is just the scalar $\nu_{\hat{x}}^i$ needed to obtain the decomposition of f. The decomposition theorem says that, in order to build a decomposition of a given continuous piecewise linear function f at a point at which the gradients of the activities are linearly independent, it suffices to know, for example, the following $|\mathcal{A}(\hat{x})| + 1$ gradients (of restrictions of f): $c^{\mathcal{A}(\hat{x})}$ and $c^{\mathcal{A}(\hat{x})-i}$, $i \in \mathcal{A}(\hat{x})$. *We do not need to use the* $2^{|\mathcal{A}(\hat{x})|}$ *gradients* corresponding to restrictions of f to each cell.

To illustrate, here is an automatic way to obtain the decomposition of function f given by (2) at the point $\hat{x} = (0,0)$. Let $\{a_i^T x - b_i\}_{i\in\mathcal{R}}$, $\mathcal{R} = \{1,2\}$, be the ridges of f, where

$$a_1 = (1,0)^T,\ a_2 = (0,1)^T$$

and $b_i = 0$, $i = 1,2$. We have $\mathcal{A}(\hat{x}) = \{1,2\}$. Take $g_{\hat{x}}$ as the gradient of f when restricted to the cell

$$\{x \in I\!R^2 : a_i^T x - b_i > 0,\ i \in \mathcal{A}(\hat{x}),\}$$

that is to say, set $g_{\hat{x}} = (-1,-1)^T$. For each $i = 1,2$, in order to obtain the scalar $\nu_{\hat{x}}^i$, one only needs to pick up two cells "separated" only by ridge i, and to compute the difference between the gradients of f restricted to each of these cells. The result is then equal to $\nu_{\hat{x}}^i a_i$, by the decomposition theorem. For example, in order to determine $\nu_{\hat{x}}^1$, we can use the cells

$$\{x \in I\!R^2 : a_1^T x - b_1 < 0 \text{ and } a_2^T x - b_2 > 0\}$$

and

$$\{x \in I\!R^2 : a_1^T x - b_1 > 0 \text{ and } a_2^T x - b_2 > 0\},$$

which yields

$$\nu_{\hat{x}}^1 a_1 = (1,-1)^T - (-1,-1)^T = (2,0)^T.$$

Since $a_1 = (1,0)^T$, we obtain $\nu_{\hat{x}}^1 = 2$. We remark that using rather the cells

$$\{x \in I\!R^2 : a_1^T x - b_1 < 0 \text{ and } a_2^T x - b_2 < 0\}$$

and

$$\{x \in I\!R^2 : a_1^T x - b_1 > 0 \text{ and } a_2^T x - b_2 < 0\},$$

we would obtain the same multiple of a_1.

In a similar manner, we obtain $\nu_{\hat{x}}^2 = 2$ and hence,

$$f(x) = g_{\hat{x}}^T x + \sum_{i\in\mathcal{A}(\hat{x})} \nu_{\hat{x}}^i \min(0, a_i^T x),$$

where $g_{\hat{x}} = (-1,-1)^T$ and $\nu_{\hat{x}}^1 = \nu_{\hat{x}}^2 = 2$.

Note that it does not appear to be necessary to automate the construction of decompositions in applications. According to our experience, decompositions (when they exist) are easily constructed in practical problems (cf. [13]).

The decomposition, $c^{\mathcal{A}(\hat{x})}$, $\{\lambda^i\}_{i\in\mathcal{A}(\hat{x})}$, constructed in the proof of the decomposition theorem, is valid for any other point $\tilde{x}$ such that $\sigma(\tilde{x}) = \sigma(\hat{x})$. Indeed, if $\tilde{x}$ is such that $\sigma(\tilde{x}) = \sigma(\hat{x})$, then any neighbourhoods of $\hat{x}$ and $\tilde{x}$, small enough, intersect exactly the same cells. For example, if the continuous piecewise linear function f is nondifferentiable at a point $\hat{x}$, and we have a decomposition g, ν^i of f at $\hat{x}$, then g, ν^i is a decomposition of f at any other point on the same segment.

In addition, when *every* point, x, of the domain of f is such that $\{a_i\}_{i\in\mathcal{A}(x)}$ are linearly independent, using an inductive argument, we can show that the $\nu_{\hat{x}}^i$'s involved in the decomposition of f at a point $\hat{x}$ are independent of $\hat{x}$. Note that from this we can prove that, assuming that the activities are linearly independent at every point $\hat{x} \in I\!R^n$, a continuous piecewise linear function f with ridges $\{a_i^T x - b_i\}_{i\in\mathcal{R}}$ has the form:

$$f(x) = b + c^T x + \sum_{i\in\mathcal{R}} \nu^i \min(0, a_i^T x - b_i), \quad \textit{for all } x \in I\!R^n,$$

for some vector $c \in I\!R^n$ and scalars b and $\{\nu^i\}_{i\in\mathcal{R}}$.

The decomposition theorem provides a sufficient condition (the linear independence of the activities) for a continuous piecewise linear function to be decomposable at a point $\hat{x}$. It would be interesting to have a condition characterizing "decomposability".

Define, for any $J \subseteq \mathcal{A}(\hat{x})$, the set

$$S^J \equiv \{x \in I\!R^n : a_i^T(x - \hat{x}) > 0, \ i \in J \text{ and } a_i^T(x - \hat{x}) < 0, \ i \in \mathcal{A}(\hat{x}) \setminus J\}.$$

Each such non-empty set corresponds to one of the cells of f intersecting any neighbourhood of $\hat{x}$. In the case where $\{a_i\}_{i\in\mathcal{A}(\hat{x})}$ are linearly dependent, we do not necessarily have $S^J \neq \emptyset$ for each $J \subseteq \mathcal{A}(\hat{x})$. Assume that f has the form (11) in a neighbourhood of $\hat{x}$, for some $c^J \in I\!R^n$ given for each $J \subseteq \mathcal{A}(\hat{x})$ such that $S^J \neq \emptyset$. Then, we can show that f is decomposable at $\hat{x}$ if and only if there exist scalars ν^i, $i \in \mathcal{A}(\hat{x})$, such that

$$\nu^i a_i = c^{J-i} - c^J, \quad \text{for all } i \in J \subseteq \mathcal{A}(\hat{x}) \tag{13}$$

such that $S^J \neq \emptyset$ and $S^{J-i} \neq \emptyset$.

This necessary and sufficient condition for f to be decomposable at a point $\hat{x}$ is however not very practical, as it involves the verification of equation (13) for *every* possible subset J of $\mathcal{A}(\hat{x})$ (such that S^J and S^{J-i} are non-empty) containing i, and this for every $i \in \mathcal{A}(\hat{x})$. We shall omit the proof.

Note however that condition (13) is satisfied at any point $\hat{x} \in I\!R^n$ for the decomposable function given by (4) (see figure 3). Also, it is clear from the form of function f_γ given by (7) that it is decomposable at any point, even when the gradient of the activities are linearly dependent. Condition (13) is straightforwardly verified at any $\hat{x} \in I\!R^n$ in this latter example. A decomposition $g_{\hat{x}}, \{\nu^i\}_{i\in\mathcal{A}(\hat{x})}$ of this function at any point $\hat{x} \in I\!R^n$ can be obtained automatically in a manner similar to the linearly independent case. If we set

$$g_{\hat{x}} \equiv \gamma c + \sum_{i\in\mathcal{A}(\hat{x})\cap E} a_i + \sum_{i\in E\setminus\mathcal{A}(\hat{x})} a_i \mathrm{sign}(a_i^T x - b_i) - \sum_{i\in I\setminus\mathcal{A}(\hat{x})} a_i s(a_i^T x - b_i), \tag{14}$$

and, for each $i \in \mathcal{A}(\hat{x})$, let J be any subset of $\mathcal{A}(\hat{x})$ such that $i \in J$, $S^J \neq \emptyset$ and $S^{J-i} \neq \emptyset$. Set, for each $i \in \mathcal{A}(\hat{x})$, ν^i to be such that

$$\nu^i a_i = c^{J-i} - c^J.$$

We thus obtain

$$\nu^i = \begin{cases} -1 & \text{if } i \in I, \\ -2 & \text{if } i \in E. \end{cases} \tag{15}$$

Hence, for all $\hat{x} \in I\!R^n$, we have:

$$f_\gamma(x) = f_\gamma(\hat{x}) + g_{\hat{x}}^T(x-\hat{x}) + \sum_{i\in\mathcal{A}(\hat{x})} \nu^i \min(0, a_i^T(x-\hat{x})),$$

for all x in some neighbourhood of $\hat{x}$ small enough, where $g_{\hat{x}}$ is defined by (14) and the scalars $\{\nu_{\hat{x}}^i\}_{i\in\mathcal{A}(\hat{x})}$ are given by (15). Note that this is an example of a function with a decomposition $g_{\hat{x}}$, $\{\nu^i\}_{i\in\mathcal{A}(\hat{x})}$ such that the scalars ν^i's are independent of $\hat{x}$. (We can show that this is a consequence of the fact that this function is decomposable at each point, as we show that the ν^i's are not dependent upon $\hat{x}$ in the case where $\{a_i\}_{i\in\mathcal{A}(\hat{x})}$ are linearly independent at each point of the domain of a function.)

3.2 Optimality Conditions

Let $\hat{x} \in I\!R^n$, $g_{\hat{x}}$ be any restricted gradient of f at $\hat{x}$, and Z be an $n \times (n - |\mathcal{A}(\hat{x})|)$ matrix whose column vectors form a basis of $\mathcal{N}(A^T(\hat{x}))$. In analogy with active set methods, where the dimension of a subproblem is reduced by considering only the space where the current activities are preserved, we call $Z^T g_{\hat{x}}$ a *reduced restricted gradient* of f at $\hat{x}$. When f is decomposable at $\hat{x}$, it is a reduced gradient of the smooth part of f. Note that Z is not unique and the reduced restricted gradient is not uniquely defined.

Let P be the orthogonal projector onto the space $\mathcal{N}(A^T(\hat{x}))$. Remark that

$$f'(\hat{x},d) = 0 \ \forall\, d \in \mathcal{N}(A^T) \ \Leftrightarrow\ g_{\hat{x}}^T d = 0 \ \forall\, d \in \mathcal{N}(A^T) \ \Leftrightarrow\ Z^T g_{\hat{x}} = \vec{0}$$

(thus the fact that $Z^T g_{\hat{x}} = \vec{0}$ is independent of which particular restricted gradient, $g_{\hat{x}}$, of f we consider) and

$$Z^T g_{\hat{x}} = \vec{0} \ \Leftrightarrow\ g_{\hat{x}} \in \mathcal{R}(A(\hat{x})) \ \Leftrightarrow\ P(g_{\hat{x}}) = \vec{0}.$$

We consider two cases:

Case 1: The reduced restricted gradient is non-null.

Consider the direction $p \equiv -P(g_{\hat{x}})$. We can easily show that along p, f decreases. Indeed, for $\alpha > 0$ small enough we have

$$\begin{aligned} f(\hat{x}+\alpha p) &= f(\hat{x}) + \alpha p^T g_{\hat{x}}, \quad \text{since } p \in \mathcal{N}(A^T), \\ &= f(\hat{x}) - \alpha \|P(g_{\hat{x}})\|^2 \\ &< f(\hat{x}). \end{aligned} \tag{16}$$

We can show that p is moreover the *steepest* descent direction among all directions in $\mathcal{N}(A^T)$.

Case 2: The reduced restricted gradient is null.

In this case, we say that $\hat{x}$ is a *dead point*. It is not possible to obtain any non-null change in the objective function f while keeping the same set of activities, as follows from equation (16), which is valid for any direction $p \in \mathcal{N}(A^T)$. Hence, we have to drop some activities if we want to obtain a descent direction. We henceforward assume that the columns of $A(\hat{x})$ are linearly independent. We return to the case of *degeneracy* (i.e. the situation where the gradients of the activities are linearly dependent) in [13]. Hence, from the decomposition theorem (theorem 1), f is decomposable at $\hat{x}$. Let $g_{\hat{x}}$, $\{\nu_{\hat{x}}^i\}_{i\in\mathcal{A}(\hat{x})}$ be a decomposition of f at

$\hat{x}$. The point $\hat{x}$ is a stationary point of the smooth part of f, and we shall either establish the optimality of f or derive a descent direction.

Consider moving from the current dead point $\hat{x}$ to a displaced point $\hat{x} + \alpha d$. We have, for $\alpha > 0$ small enough:

$$f(\hat{x} + \alpha d) = f(\hat{x}) + \alpha d^T[g_{\hat{x}} + \sum_{i \in \mathcal{A}(\hat{x}): a_i^T d < 0} \nu_{\hat{x}}^i a_i].$$

But the fact that $\hat{x}$ is a dead point implies that there exist scalars $\{u_i\}_{i \in \mathcal{A}(\hat{x})}$ such that

$$g_{\hat{x}} = \sum_{i \in \mathcal{A}(\hat{x})} u_i a_i$$

and hence,

$$f(\hat{x} + \alpha d) = f(\hat{x}) + \alpha[\sum_{i \in \mathcal{A}(\hat{x})} u_i a_i^T d + \sum_{i \in \mathcal{A}(\hat{x})} \lambda_{\hat{x}}^{i,d} a_i^T d], \qquad (17)$$

where

$$\lambda_{\hat{x}}^{i,d} = \begin{cases} \nu_{\hat{x}}^i & \text{if } a_i^T d < 0, \\ 0 & \text{otherwise.} \end{cases} \qquad (18)$$

We can show that a direction d^{k^+} (d^{k^-}) $\in \mathcal{N}(A_{-k}^T)$ dropping positively (cf. page 8 above) (negatively) activity $k \in \mathcal{A}(\hat{x})$ exists.

Hence, using the fact that $a_i^T d^{k^\pm} = 0$ for all $i \in \mathcal{A}(\hat{x}) \setminus \{k\}$, equation (17) becomes (with $d = d^{k^\pm}$):

$$f(\hat{x} + \alpha d^{k^\pm}) = f(\hat{x}) \pm \alpha(u_k + \lambda_{\hat{x}}^{k,d^{k^\pm}})|a_k^T d^{k^\pm}|, \qquad (19)$$

where '±' is a '+' when k is dropped positively and a '−' if it is dropped negatively. Thus, at a dead point, a direction dropping only activity $k \in \mathcal{A}(\hat{x})$ does not yield descent if and only if

$$0 \le u_k \le -\nu_{\hat{x}}^k.$$

The next theorem will state that in fact, it is sufficient to verify whether single-dropping directions yield descent, in order to determine if the current point is optimal.

In the same way, the optimality conditions for many nondifferentiable problems can be determined. When minimizing f_γ given by (7), one can show using (15) (and by proving the fact that one needs to consider only single-dropping directions as possible descent directions, as we shall do in the proof of theorem 2) that a feasible dead point, x^*, is optimal for the original problem if and only if $0 \le u_i^* \le 2$ for all $i \in \mathcal{A}(x^*) \cap E$ and $0 \le u_i^* \le 1$ for all $i \in \mathcal{A}(x^*) \cap I$, where the scalars $\{u_i^*\}_{i \in \mathcal{A}(x^*)}$ are the coefficients in the linear combination of the restricted gradient (given by (14)) of f at x^* in terms of the columns of $A(x^*)$ (note that the multiplier rule corresponding to an active equality constraint has the form $0 \le u_i^* \le 2$ rather than the more standard $-1 \le u_i^* \le 1$, due to the canonical form, $\sum_{i \in \mathcal{A}(\hat{x})} \nu_{\hat{x}}^i \min(0, a_i^T(x - \hat{x}))$, with which we fixed the definition of the nonsmooth part of f). The projection method for the uncapacitated facility location problem of Conn and Cornuéjols [11] determines optimality conditions and descent directions in a similar fashion, that is to say, from the observation of the behaviour of the function in each single-dropping direction.

Theorem 2 (Optimality Conditions) *Let $x^* \in I\!\!R^n$ and $f : I\!\!R^n \rightarrow I\!\!R$ be a continuous piecewise linear function with ridges $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$, where $\mathcal{R}$ is a finite index set. Assume linear independence of $\{a_i\}_{i \in \mathcal{A}(x^*)}$, and let $g_{x^*}, \{\nu_{x^*}^i\}_{i \in \mathcal{A}(x^*)}$ be the decomposition of f at x^*. The point x^* is a local minimum of f if and only if there exist scalars u_i^*, $i \in \mathcal{A}(x^*)$ such that*

1. *$g_{x^*} = \sum_{i \in \mathcal{A}(x^*)} u_i^* a_i$ (or, equivalently, the reduced restricted gradient of f at x^* is null, i.e. x^* is a dead point) and*
2. *$0 \leq u_i^* \leq -\nu_{x^*}^i$, for all $i \in \mathcal{A}(x^*)$.*

A simple interpretation of the condition $0 \leq u_i^* \leq -\nu_{x^*}^i$ when x^* is a dead point, is that the directional derivative of f in the single-dropping directions $d^{i^\pm}$ (dropping $i \in \mathcal{A}(x^*)$) is non-negative.

Proof: The necessity of the optimality conditions follows from the earlier discussion. Note that if x^* is a dead point, then the coefficients $\{u_i^*\}_{i \in \mathcal{A}(x^*)}$ exist and are uniquely defined, by the assumption of the linear independence of $\{a_i\}_{i \in \mathcal{A}(x^*)}$. If we have $u_k^* < 0$ ($u_k^* > \nu_{x^*}^k$) for some $k \in \mathcal{A}(x^*)$, then a direction dropping activity k positively (negatively) is a descent direction.

In order to prove the sufficiency of the optimality conditions, suppose that the two conditions hold. Consider $d \in I\!\!R^n$, an arbitrary direction. Recall from equation (17), which was valid for an arbitrary d at a dead point, that

$$f(x^* + \alpha d) = f(x^*) + \alpha \sum_{i \in \mathcal{A}(x^*)} (u_i^* + \lambda_{x^*}^{i,d}) a_i^T d, \tag{20}$$

where $\lambda_{x^*}^{i,d}$ is defined by (18). Hence,

$$f(x^* + \alpha d) \geq f(x^*),$$

by hypothesis, and thus d is not a descent direction. □

3.3 Algorithm

We now present an algorithm for minimizing a continuous piecewise linear function, $f : I\!\!R^n \rightarrow I\!\!R$. We assume that f is decomposable at each iterate and at each breakpoint encountered in the line search. Moreover, we assume that at every iterate which is a dead point, the gradients of the activities are linearly independent.

Let $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$ be the ridges of f, where $\mathcal{R}$ is a finite index set, and $g_{x^k}, \{\nu_{x^k}^i\}_{i \in \mathcal{A}(x^k)}$ be the decomposition of f at x^k.

Continuous Piecewise Linear Minimization Algorithm

Step 1: Choose any $x^1 \in I\!\!R^n$ and set $k \leftarrow 1$.

Step 2: Identify the activities, $\mathcal{A}(x^k)$, and compute $d^k \equiv -P(g_{x^k})$, the projection of the restricted gradient onto the space orthogonal to the gradients of the activities. If $d^k \neq \vec{0}$ then go to step 6.

(Now x^k is a dead point. Compute a single-dropping descent direction or establish optimality.)

Step 3: Compute $\{u_i\}_{i\in\mathcal{A}(x^k)}$, the coefficients of $\{a_i\}_{i\in\mathcal{A}(x^k)}$ in the linear combination of g_{x^k} in terms of the columns of $A(x^k)$.

Step 4: If $u_i < 0$ or $u_i > -\nu^i_{x^k}$, for some $i \in \mathcal{A}(x^k)$ (violated optimality condition), then go to step 6.

Otherwise, stop: x^k is a local minimum of f.

Step 5: (Drop activity i)

Redefine $d^k = P_{-i}(a_i)$, if the violated inequality found corresponds to $u_i \geq 0$, otherwise $d^k = -P_{-i}(a_i)$ and $u_i \leq -\nu^i_{x^k}$, where P_{-i} is the orthogonal projector onto the space orthogonal to the gradients of all the activities but activity i.

Step 6: (Line search)

Determine the step size α^k by solving $\min_{\alpha>0} f(x^k + \alpha d^k)$.

This line search can be done from x^k, moving from one breakpoint of f to the next, in the direction d^k, until either we establish unboundedness of the objective function or the value of f starts increasing.

Step 7: Update $x^{k+1} = x^k + \alpha^k d^k$, $k \leftarrow k + 1$ and go to step 2.

Remarks:

- In practice, the orthogonal projectors P and P_{-i} are computed with a suitable factorization and/or update of the matrix of activities, $A(x^k)$.
- In order to compute the coefficients $\{u_i\}_{i\in\mathcal{A}(x^k)}$ in step 3, one needs to solve a linear system which is possibly overdetermined but always feasible, since $g_{x^k} \in \mathcal{R}(A(x^k))$ at a dead point x^k. This can easily be done by solving the least squares problem:

$$\min_u \|g_{x^k} - A(x^k)u\|$$

(where the vector u is indexed by $\mathcal{A}(x^k)$), since we already have a factorization of the matrix of activities.

- In step 5, when $u_i < 0$ or $u_i > -\nu^i_{x^k}$, we can show that $d^k = \pm P_{-i}(a_i)$ is the *steepest* descent direction in the space $\mathcal{N}(A^T_{-i})$.
- In step 6, we update the directional derivative of the objective function in the direction d^k from one breakpoint to the other. For example, if d^k was obtained from step 2, then the directional derivative of f_{γ^k} at x^k in the direction d^k is

$$\begin{aligned} f'_{\gamma^k}(x^k; d^k) &= g^T_{x^k} d^k = -g^T_{x^k} P(g_{x^k}) \\ &= -g^T_{x^k} P^T P(g_{x^k}) = -\|d^k\|^2. \end{aligned} \tag{21}$$

In the case where the descent direction is obtained from step 5, dropping activity i, we have

$$f'_{\gamma^k}(x^k; d^k) = (u_i + \lambda^{i,d^k}_{x^k}) a_i^T d^k$$

(where $\lambda^{i,d^k}_{x^k}$ is defined by (18)). If we then encounter a breakpoint $\bar{x}$ at which f_{γ^k} is decomposable, with decomposition $g_{\bar{x}}$, $\{\nu^i_{\bar{x}}\}_{i\in\mathcal{A}(\bar{x})}$, in the line search when "crossing" exactly one ridge $j \in \mathcal{A}(\bar{x})$ (i.e $\mathcal{A}(\bar{x}) \setminus \mathcal{A}(x^k) = \{j\}$), then the directional derivative becomes

$$\begin{aligned}
f'_{\gamma^k}(\bar{x};d^k) &= (g_{\bar{x}} + \lambda_{\bar{x}}^{j,d^k} a_j)^T d^k \\
&= (g_{\bar{x}} + \lambda_{\bar{x}}^{j,d^k} a_j)^T d^k + \lambda_{\bar{x}}^{j,-d^k} a_j^T d^k - \lambda_{\bar{x}}^{j,-d^k} a_j^T d^k \\
&= (g_{\bar{x}} + \lambda_{\bar{x}}^{j,-d^k} a_j)^T d^k + \lambda_{\bar{x}}^{j,d^k} a_j^T d^k - \lambda_{\bar{x}}^{j,-d^k} a_j^T d^k \\
&= -(g_{\bar{x}} + \lambda_{\bar{x}}^{j,-d^k} a_j)^T (-d^k) + \lambda_{\bar{x}}^{j,d^k} a_j^T d^k - \lambda_{\bar{x}}^{j,-d^k} a_j^T d^k \\
&= -f'_{\gamma^k}(\bar{x};-d^k) + \lambda_{\bar{x}}^{j,d^k} a_j^T d^k - \lambda_{\bar{x}}^{j,-d^k} a_j^T d^k \\
&= f'_{\gamma^k}(x^k;d^k) + \lambda_{\bar{x}}^{j,d^k} a_j^T d^k - \lambda_{\bar{x}}^{j,-d^k} a_j^T d^k \qquad (22) \\
&= f'_{\gamma^k}(x^k;d^k) - \nu_{\bar{x}}^j |a_j^T d^k|.
\end{aligned}$$

Equality (22) follows immediately since f_{γ^k} is piecewise linear. If at the first breakpoint, $\bar{x}$, we cross a *set* $\mathcal{J}$ of many ridges (i.e $\mathcal{A}(\bar{x}) \setminus \mathcal{A}(x^k) = \mathcal{J}$), then the update is done similarly:

$$f'_{\gamma^k}(\bar{x};d^k) = f'_{\gamma^k}(x^k;d^k) - \sum_{j \in \mathcal{J}} \nu_{\bar{x}}^j |a_j^T d^k|. \qquad (23)$$

In order to know whether the value of f_{γ^k} starts increasing beyond breakpoint $\bar{x}$, we check the sign of $f'_{\gamma^k}(\bar{x};d^k)$, the directional derivative at $\bar{x}$ in the direction d^k.

Note that a *full* ordering of the step sizes corresponding to each breakpoint along the line search is not necessary. We need to have access to these step sizes one at a time, in order of increasing size, only until the value of f_{γ^k} starts increasing. We use the "smallest-in/first-out" mechanism of a *heap* (e.g. see [28]).

When studying the uncapacitated facility location problem, Conn and Cornuéjols [11] had to minimize the continuous piecewise linear convex objective function

$$f(x) = \sum_{i \in I} x_i + \sum_{j \in J} [\sum_{i \in I} (c_{ij} - x_i)^+ - f_j]^+,$$

where $a^+ \equiv \max(0,a)$ and c_{ij} and f_j are constants. We are here in the presence of "nested ridges" (the derivative of f is not defined over $\{x : r(x) \equiv \sum_{i \in I}(c_{ij} - x_i)^+ - f_j = 0\}$, where the function r has ridges, namely: $c_{ij} - x_i$, $i \in I$). They derived optimality conditions which are specific to this problem in order to construct an algorithm. To illustrate how this problem fits our general framework, consider the following simple instance: $f(x,y) = [(-x)^+ + (-y)^+ - 1]^+$. Figure 4 shows the values f takes over the different cells. Figure 5 gives the decomposition, $g_{\bar{x}}$, $\{\nu_{\bar{x}}^i\}_{i \in \mathcal{A}(\bar{x})}$, of f at every point where the function is decomposable (i.e. every point of $\mathbb{R}^2 \setminus \{(0,-1)^T, (-1,0)^T\}$). The circles in figure 5 contain the different values the restricted gradient, $g_{\bar{x}}^T$, takes over the different subdomains where f is decomposable. In a region over which f is differentiable, $g_{\bar{x}}$ is just the gradient of the restriction of f to that region. The squares contain the ν^i values for each segment of ridge i (with $\nu^i = 0$ for the dashed line part of the ridge). Thus, apart from the two points $(0,-1)^T$ and $(-1,0)^T$ (at which the gradients of the activities are linearly dependent), we have a decomposition of f at any point of the plane. At the two degenerate points $(0,-1)^T$ and $(-1,0)^T$, one can easily show (we shall prove the "non-decomposability" of a function in [13]) that f is not decomposable (this can be seen also using the characterization of decomposable functions given by (13)). Thus, special attention will have to be paid at such singular points.

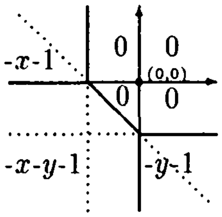

Figure 4: Simple instance of objective function for the uncapacitated facility location problem

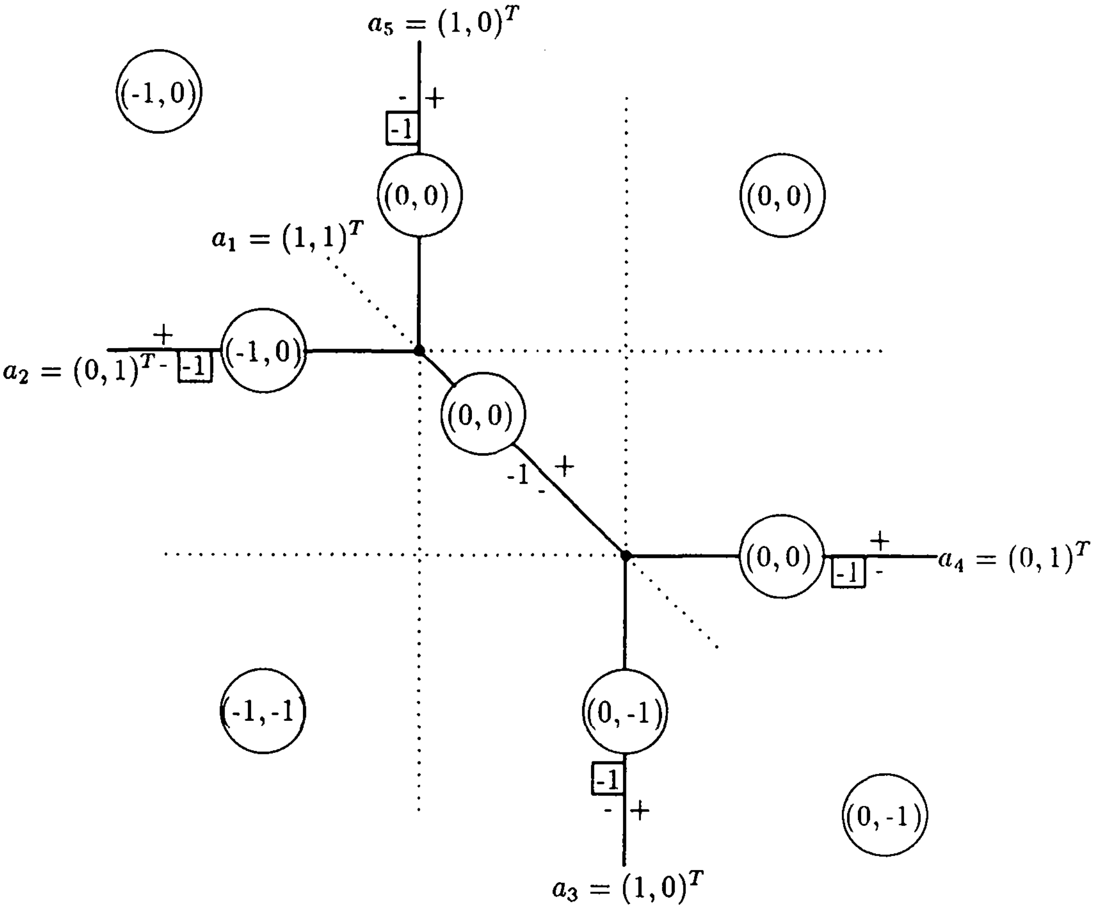

Figure 5: Example of a decomposition

3.4 Finite Step Convergence

Theorem 3 (Convergence) *Let $f : I\!R^n \rightarrow I\!R$ be a continuous piecewise linear function with ridges $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$, where $\mathcal{R}$ is a finite index set. Assume that*

1. *$\{a_i\}_{i \in \mathcal{A}(x^k)}$ are linearly independent at each iterate, x^k, encountered in the course of the algorithm, and*
2. *f is decomposable at each breakpoint encountered along the line searches.*

Then the continuous piecewise linear minimization algorithm converges globally (i.e. from any starting point) in a finite number of iterations.

Proof: Starting at any point, the algorithm generates a sequence of points, each of which is obtained from its predecessor using step 7. The step size α^k in step 6 is chosen so that a ridge that was not active at x^k is now active at x^{k+1}. If iterate x^k is not a dead point, then the direction vector d^k for generating x^{k+1}, the succeeding point, will be chosen at step 2. This will guarantee that all ridges which are active at x^k will be active at x^{k+1}. Hence, in this case

$$\mathcal{A}(x^k) \subset \mathcal{A}(x^{k+1}).$$

Thus, under the assumption of the linear independence of the gradients of the activities, the sequence step 2–step 8–step 10 cannot be executed consecutively more than n times.

Again by the linear independence hypothesis, when x^k is a dead point, the descent direction will drop only *one* activity and hence

$$|\mathcal{A}(x^k)| \leq |\mathcal{A}(x^{k+1})|$$

(at x^{k+1} there is exactly one activity fewer than at x^k, but at least one new activity was hit). By definition of a dead point, we have, for α small enough:

$$f(x^k + \alpha p) = f(x^k), \quad \text{for all } p \in \mathcal{N}(A(x^k)^T)$$

and hence, $f(x)$ is constant throughout

$$\{x : \sigma(x) = \sigma(x^k)\}.$$

The sequence of points generated by the algorithm determine a monotonic decreasing sequence of values of f. This means that one cannot return to a point x such that $\sigma(x) = \sigma(x^k)$. Since there is only a finite number of different vectors $\sigma \in 3^{|\mathcal{R}|}$, we may conclude that only a finite number of dead points can be generated. On the other hand, if the original problem is not unbounded, then by the optimality conditions theorem (theorem 2) some dead point is optimal. The algorithm will not terminate until an optimal dead point has been found or a conclusion of unboundedness is reached. □

4 Discontinuous Piecewise Linear Optimization

We now consider the case where, again, f is piecewise linear (with ridges $\{a_i^T - b_i\}_{i \in \mathcal{R}}$, where $\mathcal{R}$ is a finite index set) but with possibly discontinuities across some ridges. We shall term such ridges: *faults* and $\mathcal{F}(\hat{x})$ will denote the faults which are active at $\hat{x}$.

A (local) minimum does not always exist in the discontinuous case. Consider for example the following univariate function, having $x = 0$ as a fault:

$$f(x) = \begin{cases} x+1 & \text{if } x \geq 0 \\ -x & \text{otherwise.} \end{cases} \tag{24}$$

Hence, we shall rather look for a local *infimum*. In order to find such a local infimum of a function f having some faults, we shall simply generalize the algorithm for the continuous problem by implicitly considering any discontinuity or *jump* across a fault i in f as the limiting case of a continuous situation.

Since we are looking for a local *infimum* of a given function f, it is equivalent to work rather with the function $\underline{f}$ defined by

$$\underline{f}(x) \equiv \liminf_{\bar{x} \to x} f(\bar{x}).$$

Thus we need only to look for a local *minimum* of $\underline{f}$. This convention will simplify the exposition. Without loss of generality, we shall henceforward only consider functions f such that $f(x) = \underline{f}(x)$ (in other words, we consider the *lower semicontinuous envelope* of f).

4.1 Soaring Directions and Faults

The algorithm will be essentially the same as in the continuous case except that we consider dropping an active fault from a dead point, x, only if we do so along a direction d such that

$$\lim_{\delta \to 0^+} f(x + \delta d) = f(x)$$

(i.e. as $\delta > 0$ is small, the value of f does not *jump up* from x to $x + \delta d$). Thus, virtually only step 4 must be adapted from the continuous problem algorithm in order to solve the discontinuous case. Note that in step 4, there is no single-dropping direction that corresponds to "jumping down" since by our convention ($f = \underline{f}$) we have

$$f(x^k) \equiv \liminf_{x \to x^k} f(x).$$

To make more rigorous the intuitive concept of directions jumping up or down, we define the set of *soaring directions* from a point $\hat{x}$ to be:

$$S(\hat{x}) \equiv \{d \in I\!\!R^n : \exists \epsilon > 0, \bar{\delta} > 0 \text{ such that } \forall\, 0 < \delta < \bar{\delta},\ f(\hat{x} + \delta d) - f(\hat{x}) > \epsilon\}.$$

If we define, for a non-degenerate point $\hat{x}$,

$$\mathcal{S}^+(\hat{x}) \equiv \{i \in \mathcal{A}(\hat{x}) : \text{ if } d^{i^+} \in \mathcal{N}(A_{-i}^T) \text{ and } a_i^T d^{i^+} > 0 \text{ then } d^{i^+} \in S(\hat{x})\}$$

and

$$\mathcal{S}^-(\hat{x}) \equiv \{i \in \mathcal{A}(\hat{x}) : \text{ if } d^{i^-} \in \mathcal{N}(A_{-i}^T) \text{ and } a_i^T d^{i^-} < 0 \text{ then } d^{i^-} \in S(\hat{x})\},$$

then the set of soaring single-dropping directions from $\hat{x}$ are simply the directions dropping an activity $i \in \mathcal{S}^+(\hat{x})$ positively and the directions dropping an $i \in \mathcal{S}^-(\hat{x})$ negatively.

A fault can now be defined more rigorously: a *positive (negative) fault of f at a point $\hat{x}$* is a ridge $i \in \mathcal{R}$ such that for any neighbourhood, $B(\hat{x})$, of $\hat{x}$, there exists a nondegenerate point $x' \in B(\hat{x})$ with $i \in \mathcal{S}^+(x')$ (with $i \in \mathcal{S}^-(x')$). The set of all positive (negative) faults at $\hat{x}$ is denoted by $\mathcal{F}^+(\hat{x})$ ($\mathcal{F}^-(\hat{x})$). The set of *faults of f at a point* $\hat{x}$ is denoted by

$$\mathcal{F}(\hat{x}) \equiv \mathcal{F}^+(\hat{x}) \cup \mathcal{F}^-(\hat{x}).$$

[Note that the definition of a fault $i \in \mathcal{F}(\hat{x})$ is described via *single*-dropping directions, dropping activity i, in order to ensure that the jump is indeed "caused" by activity i (and is not due to some other fault which would also be active at $\hat{x}$). Also, $\mathcal{F}^+$ and $\mathcal{F}^-$ are defined in terms of soaring directions in a *neighbourhood* of x_c and not at x_c alone. Moreover, we need to refer to some point x' in a neighbourhood of $\hat{x}$, so that the definition makes sense at a degenerate point $\hat{x}$—recall that a single-dropping direction does not necessarily exist from a degenerate point].

We would like to modify the continuous problem algorithm in such a way that, at a non-degenerate dead point, x^k, we do not need to verify the optimality conditions corresponding to soaring single-dropping directions ($u_i \geq 0$, $i \in \mathcal{S}^+(x^k)$ and $u_i \leq -\nu^i$, $i \in \mathcal{S}^-(x^k)$), so that we would never consider such single-dropping directions in order to establish whether x^k is optimal. This is reasonable since we are looking for a *local* minimum. The line search step (step 6) will be modified similarly: when we encounter a breakpoint $\bar{x}$ on a fault along a direction $d \in S(\bar{x})$ (jump up), we stop and if d is such that $-d \in S(\bar{x})$, (jump down), we carry on to the next breakpoint, and update properly the directional derivative along d. We first need to define a decomposition of a discontinuous function at a point on a fault.

4.2 Decomposition

In the non-degenerate case, we can always find a decomposition of a continuous piecewise linear function f such that the change of the nonsmooth part of f in a single-dropping direction, d, can be written as the dot product of d with a multiple of the gradient of the activity to be dropped. The proof of this fact was based on the continuity of f, as we saw in the proof of the decomposition theorem (theorem 1). This is what enabled us to isolate the effect of each activity, yielding an optimality condition corresponding to each of the possible single-dropping directions. Apart from the fact that it relies on the scalars $\{u_i\}_{i \in \mathcal{A}}$, each of these conditions was independent of the activities other than the one being dropped. Nevertheless, in the discontinuous case, when $a_i^T x - b_i$ is a fault at a non-degenerate point, $\hat{x}$, there is in general also a way to define a decomposition, g, $\{\nu^i\}_{i \in \mathcal{A}}$, of f at $\hat{x}$ so that when we drop $i \notin \mathcal{S}^+(\hat{x})$ positively or $i \notin \mathcal{S}^-(\hat{x})$ negatively (i.e. in a non-soaring single-dropping direction), the change of the nonsmooth part of f can be expressed as the dot product of the single-dropping direction with a multiple of a_i.

Note that if $i \in \mathcal{R}$ is a positive fault at some point $\hat{x}$, and a negative fault at some other point $\tilde{x}$ on the same segment (i.e. $\sigma(\hat{x}) = \sigma(\tilde{x})$), R, then, in practice, we shall consider R

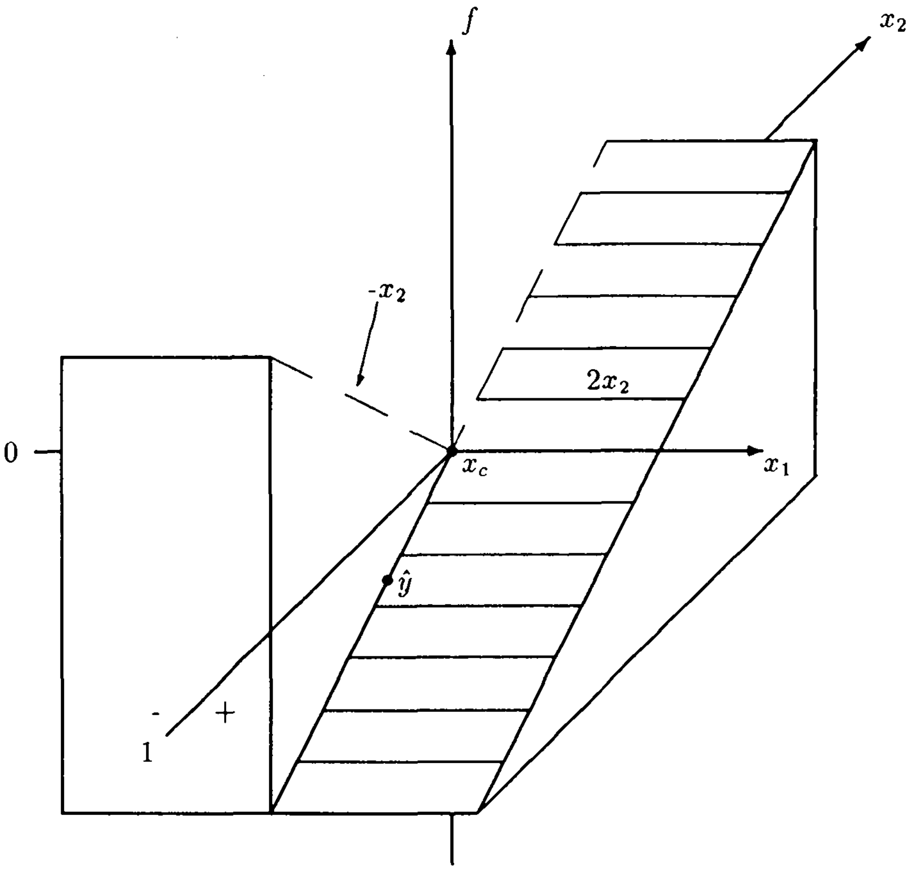

Figure 6: Graph of a function having a contact point x_c

as divided further into two segments. To do so, we can introduce another (*artificial*) ridge, $a_0^T x - b_0$, the sign of whose residuals will be used in order to know to which segment belongs an iterate $x^k \in R$. Note that one has to be careful at a "contact" point $x_c \in I\!R$ (defined below) such that $a_0^T x - b_0 = 0$. At x_c, contrary to at some other points of R, we can drop activity i *both* positively and negatively.

The function $f : I\!R^2 \to I\!R$, given by

$$f(x) = \begin{cases} 2x_2 & \text{if } x_1 > 0 \text{ or } (x_1 = 0 \text{ and } x_2 \leq 0), \\ -x_2 & \text{otherwise,} \end{cases} \tag{25}$$

illustrates well the situation. Figure 6 shows the graph of f in a neighbourhood of $x_c \equiv (0,0)^T$. We introduce the artificial ridge

$$a_0^T x - b_0 \equiv x_2 = 0,$$

so that the fault $x_1 = 0$ is partitioned into three segments according to whether $a_0^T x - b_0$ is positive, negative or zero. The point x_c is a contact point with respect to the fault $x_1 = 0$.

(The choice of a particular vector a_0 for an artificial ridge is rather arbitrary but necessary. Note however that for the contact points present in the applications of [13], we do not need to introduce such arbitrary ridges, as we can use existing ridges of the objective function to characterize the points of a fault from which the fault could be dropped both positively and negatively.)

Formally, we define $x_c \in I\!R^n$ to be a *contact point of f with respect to $i \in \mathcal{A}(x_c)$*, when $i \in \mathcal{F}(x_c)$ such that either

1. $i \in \mathcal{F}^+(x_c) \cap \mathcal{F}^-(x_c)$, or
2. there exist $\sigma^+, \sigma^- \in 3^{|\mathcal{R}|}$ such that $\sigma_i^+ = 1$, $\sigma_i^- = -1$ and

$$\lim_{x \to x_c, \sigma_i(x) = \sigma^+} f(x) = \lim_{x \to x_c, \sigma_i(x) = \sigma^-} f(x)$$

(continuity when crossing ridge i, which is a fault, at x_c).

We term $\hat{x} \in I\!R^n$ a *non-contact point* of f when $\hat{x}$ is not a contact point of f with respect to any $i \in \mathcal{A}(x_c)$.

Note that the fault $x_1 = 0$ and the point $x_c = (0,0)^T$ satisfy both conditions 1 and 2 in the above definition of a contact point for the function f defined by (25). They however satisfy only condition 1 for the function $f : I\!R^2 \to I\!R$, defined by

$$f(x) = \begin{cases} 1 & \text{if } x_1 \geq 0 \text{ and } x_2 \geq 0, \\ 2 & \text{if } x_1 \leq 0 \text{ and } x_2 \leq 0 \text{ and } x \neq (0,0)^T, \\ 3 & \text{if } x_1 > 0 \text{ and } x_2 < 0, \\ 4 & \text{otherwise.} \end{cases}$$

For the function $f : I\!R^2 \to I\!R$ given by

$$f(x) = \begin{cases} -x_2 & \text{if } x_1 > 0 \text{ and } x_2 < 0, \\ 0 & \text{otherwise,} \end{cases}$$

they satisfy only condition 2 ($\mathcal{F}^-(x_c)$ is empty).

For the function f defined by (25), we have the following trivial "decomposition" (we shall define it shortly for the discontinuous case) at the point $\hat{y}$ on the fault $x_1 = 0$ (see figure 6) such that $a_0^T \hat{y} - b_0$ is negative:

$$f(x) = f(\hat{y}) + g_{\hat{y}}^T (x - \hat{y}) + \nu_{\hat{y}}^1 \min(0, a_1^T (x - \hat{y})), \tag{26}$$

for all $x \notin S(\hat{y}) + \hat{y}$, where

$$a_1^T = (1,0), \; g_{\hat{y}} = (0,2)^T \text{ and } \nu_{\hat{y}}^1 = 0.$$

Since $\{x \in I\!R^2 : x \notin S(\hat{y}) + \hat{y}\}$ is simply the set of points x satisfying: $a_1^T x - b_1 \geq 0$, the scalar $\nu_{\hat{y}}^1$ is clearly meaningless and superfluous (since in (26), $\min(0, a_1^T(x - \hat{y}))$ is always zero for such points). This "decomposition" simply gives the behaviour of f (equation (26)) for every point x in the neighbourhood of $\hat{y}$ such that $x - \hat{y}$ is not a soaring direction from $\hat{y}$. An algorithm similar to the one introduced in the continuous case, but which does not consider *soaring* single-dropping directions, will encounter no difficulty with the discontinuity in f at any non-contact point (i.e. for (25) at any point other than x_c). We shall see that the fact that the decomposition is undefined for the x's such that $x - \hat{x}$ is a soaring direction from a point $\hat{x}$, will not affect our method (we never consider the soaring single-dropping directions). The ridge $x_1 = 0$ is a negative fault of f (figure 6) at the point $\hat{y}$. It can only possibly be dropped positively (negatively we would "jump up" the fault). Thus, if $\hat{y}$ were a dead point, we would only verify whether the optimality condition $u_1 \geq 0$ is satisfied in order to see if dropping the fault $x_1 = 0$ yields descent. Note that at the contact point, $x_c = (0,0)$, the fault could be dropped either positively or negatively as none of the two single-dropping directions, $(1,0)^T$ and $(-1,0)^T$, from x_c are a soaring direction. At a point $\hat{z}$ on the fault such that $a_0^T \hat{z} - b_0$ is positive, we can also "decompose" f exactly as in (26), with, this time, $g_{\hat{z}} = (0,-1)^T$, if we defined ridge 1 to be such that $a_1^T = (-1,0)$.

We assume that a (possibly) discontinuous piecewise linear function, f, is given under the same form as in the continuous case, except that here, we assume moreover that at any point $\hat{x} \in \mathbb{R}^n$, we can obtain $\mathcal{F}^+(\hat{x})$ and $\mathcal{F}^-(\hat{x})$, the set of the positive faults and the set of the negative faults of f at $\hat{x}$.

Without loss of generality, we shall henceforward assume that at a *given non-contact point* $\hat{x}$, all the faults of f at $\hat{x}$ are negative faults at $\hat{x}$ (otherwise we could replace a_i by $-a_i$ and b_i by $-b_i$), that is to say $\mathcal{F}(\hat{x}) = \mathcal{F}^-(\hat{x})$.

We now redefine the notion of decomposition at a point $\hat{x}$ to take into account the possibility of having $\hat{x}$ on a fault.

Definition 2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a (possibly discontinuous) piecewise linear function with ridges $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$, where $\mathcal{R}$ is a finite index set, and let $\hat{x} \in \mathbb{R}^n$. Let $g_{\hat{x}} \in \mathbb{R}^n$ and $\psi_{\hat{x}}$ be a function defined on $\mathbb{R}^n$ such that we have:*

$$f(x) = f(\hat{x}) + g_{\hat{x}}^T (x - \hat{x}) + \psi_{\hat{x}}(x), \tag{27}$$

for all $x \in B(\hat{x})$ such that $a_i^T(x - \hat{x}) \geq 0$, $i \in \mathcal{F}(\hat{x})$, for some neighbourhood, $B(\hat{x})$, of $\hat{x}$, where

$$\psi_{\hat{x}}(x) = \sum_{i \in \mathcal{A}(\hat{x}) \setminus \mathcal{F}(\hat{x})} \nu_{\hat{x}}^i \min(0, a_i^T(x - \hat{x}))$$

for some scalars $\{\nu_{\hat{x}}^i\}_{i \in \mathcal{A}(\hat{x}) \setminus \mathcal{F}(\hat{x})}$.

We say that $g_{\hat{x}}, \{\nu_{\hat{x}}^i\}_{i \in \mathcal{A}(\hat{x}) \setminus \mathcal{F}(\hat{x})}$ is a decomposition *of f* (into a smooth function and a sum of functions having a single ridge) *at $\hat{x}$.*

Based on this definition of decomposition of a discontinuous piecewise linear function f, we say again that f is *decomposable* at $\hat{x}$ when there exists a decomposition of f at $\hat{x}$.

To illustrate, consider the discontinuous function $f : \mathbb{R}^2 \to \mathbb{R}$:

$$f(x) = \min(0, x_1 - x_2) + |x_2| + \begin{cases} x_2 & \text{if } x_1 \geq 0, \\ -x_2 + 1 & \text{otherwise,} \end{cases}$$

which has the set of ridges $\{a_i^T x - b_i\}_{i\in\mathcal{R}}$, where $\mathcal{R} \equiv \{1,2,3\}$,

$$a_1 = (1,-1)^T, \ \ a_2 = (0,1)^T, \ \ a_3 = (1,0)^T, \ \ b_1 = b_2 = b_3 = 0.$$

Here is a decomposition of this function at $\hat{x} \equiv (0,0)^T$ (note that $\mathcal{F}(\hat{x}) = \mathcal{F}^-(\hat{x}) = \{3\}$):

$$f(x) = g_{\hat{x}}^T x + \sum_{i\in\mathcal{A}(\hat{x})\setminus\mathcal{F}(\hat{x})} \nu^i \min(0, a_i^T x)$$

for all x in some small enough neighbourhood of $\hat{x}$ and such that $a_i^T(x-\hat{x}) \geq 0, \ i \in \mathcal{F}(\hat{x})$ (i.e. such that $x_1 \geq 0$), where

$$g_{\hat{x}} = (0,2)^T \ \ \nu^1 = 1 \ \text{ and } \ \nu^2 = -2.$$

The decomposition theorem (theorem 1) and proposition 1 proved in the continuous case still hold at non-contact points when referring to definition 2 (the case where $\hat{x}$ is a contact point is discussed in [13], where all aspects of degeneracy and contact points are considered):

Theorem 4 (Decomposition—Discontinuous Case) *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a (possibly discontinuous) piecewise linear function with ridges $\{a_i^T x - b_i\}_{i\in\mathcal{R}}$, where $\mathcal{R}$ is a finite index set, and let $\hat{x} \in \mathbb{R}^n$ be such that $\mathcal{F}^+(\hat{x}) \cap \mathcal{F}^-(\hat{x}) = \emptyset$. Without loss of generality, assume that $\mathcal{F}(\hat{x}) = \mathcal{F}^-(\hat{x})$. If $\{a_i\}_{i\in\mathcal{A}(\hat{x})}$, the gradients of the ridges of f which are active at $\hat{x}$, are linearly independent, then f is decomposable at $\hat{x}$ and the decomposition is unique.*

Proof: The proof goes exactly as the proof of the continuous case (theorem 1), except for the fact that we are now dealing with the points x in a neighbourhood of $\hat{x}$ such that $a_i^T(x-\hat{x}) \geq 0, \ i \in \mathcal{F}(\hat{x})$, and the $c^J \in \mathbb{R}^n$ are given for each $J \subseteq \mathcal{A}(\hat{x})$ such that $\mathcal{F}(\hat{x}) \subseteq J$. We use the continuity of f over

$$\{x \in N(\hat{x}) : a_i^T(x-\hat{x}) > 0, i \in \mathcal{F}(\hat{x})\}$$

to obtain: for $I, J \subseteq \{1,2,\ldots,m\}$ such that $\mathcal{F}(\hat{x}) \subseteq I \cap J$,

$$d_i^I = d_i^J \ \text{ whenever } i \in I \cap J \text{ or } i \in \{1,2,\ldots,m\} \setminus (I \cup J) \text{ or } m < i \leq n,$$

which is true if and only if for $I \subseteq \{1,2,\ldots,m\}$ such that $\mathcal{F}(\hat{x}) \subseteq I$, and $1 \leq i \leq n$:

$$d_i^I = \begin{cases} d_i^{\{1,2,\ldots,m\}} & \text{if } i \in I, \\ d_i^{\mathcal{F}(\hat{x})} & \text{otherwise.} \end{cases}$$

Hence, for all $y \in \mathbb{R}^n$ such that $y_i \geq 0, \ i \in \mathcal{F}(\hat{x})$, we have

$$h(y) = \sum_{i=1}^{n} d_i y_i,$$

where

$$d_i = \begin{cases} d_i^{\{1,2,\ldots,m\}} & \text{if } y_i \geq 0, \\ d_i^{\mathcal{F}(\hat{x})} & \text{if } y_i < 0. \end{cases}$$

Thus, as in the proof of theorem 1, we obtain, with

$$\nu^i \equiv d_i^{\mathcal{F}(\hat{x})} - d_i^{\{1,2,\ldots,m\}},\ i \in \{1,2,\ldots,m\} \setminus \mathcal{F}(\hat{x}) :$$

$$f(x) = x^T c^{\{1,2,\ldots,m\}} + \sum_{i \in \{1,2,\ldots,m\} \setminus \mathcal{F}(\hat{x})} \nu^i \min(0, a_i^T x), \tag{28}$$

for all x in a neighbourhood of $\hat{x}$ such that $a_i^T(x - \hat{x}) \geq 0$, $i \in \mathcal{F}(\hat{x})$.

Hence, (28) is a decomposition of f at $\hat{x}$.

The proof of the uniqueness follows exactly as in of theorem 1. □

Proposition 2 *Let $f : I\!R^n \to I\!R$ be a (possibly discontinuous) piecewise linear function with ridges $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$, where $\mathcal{R}$ is a finite index set, and let $\hat{x} \in I\!R^n$ be such that $\mathcal{F}^+(\hat{x}) \cap \mathcal{F}^-(\hat{x}) = \emptyset$. Without loss of generality, assume that $\mathcal{F}(\hat{x}) = \mathcal{F}^-(\hat{x})$. Assume that $\{a_i\}_{i \in \mathcal{A}(\hat{x})}$, the gradients of the ridges of f which are active at $\hat{x}$, are linearly independent and that, without loss of generality, f has the following form for all x such that $a_i^T(x - \hat{x}) \geq 0$, $i \in \mathcal{F}(\hat{x})$, in a neighbourhood of $\hat{x}$:*

$$f(x) = f(\hat{x}) + (x - \hat{x})^T c^J$$

whenever

$$\begin{array}{rrcl} & a_i^T(x - \hat{x}) & \geq & 0,\ i \in J \\ \text{and} & a_i^T(x - \hat{x}) & < & 0,\ i \in \mathcal{A}(\hat{x}) \setminus J, \end{array}$$

for some $c^J \in I\!R^n$ given for each $J \subseteq \mathcal{A}(\hat{x})$ such that $\mathcal{F}(\hat{x}) \subseteq J$.

Then, $c^{\mathcal{A}(\hat{x})}, \{\lambda^i\}_{i \in \mathcal{A}(\hat{x}) \setminus \mathcal{F}(\hat{x})}$ is the decomposition of f at $\hat{x}$, where each of the scalars $\{\lambda^i\}_{i \in \mathcal{A}(\hat{x}) \setminus \mathcal{F}(\hat{x})}$ is such that

$$\lambda^i a_i = c^{J_i - i} - c^{J_i},$$

for any *given $J_i \subseteq \mathcal{A}(\hat{x})$ such that $i \in J_i$ and $\mathcal{F}(\hat{x}) \subseteq J_i$.*

Proof: We adapt straightforwardly the proof of proposition 1. Since f is decomposable at $\hat{x}$, there exists some decomposition $g, \{\nu^i\}_{i \in \mathcal{A}(\hat{x}) \setminus \mathcal{F}(\hat{x})}$ of f at $\hat{x}$. Hence, for any $\mathcal{F}(\hat{x}) \subseteq J \subseteq \mathcal{A}(\hat{x})$ we have

$$c^J = g + \sum_{k \in \mathcal{A}(\hat{x}) \setminus J} \nu^k a_k,$$

since

$$f(x) = f(\hat{x}) + g_{\hat{x}}^T(x - \hat{x}) + \sum_{i \in \mathcal{A}(\hat{x}) \setminus \mathcal{F}(\hat{x})} \nu^i \min(0, a_i^T(x - \hat{x}))$$

for all $x \in B(\hat{x})$ such that $a_i^T(x - \hat{x}) \geq 0$, $i \in \mathcal{F}(\hat{x})$, for some neighbourhood, $B(\hat{x})$, of $\hat{x}$. Thus, we can show that this implies $c^{J-i} - c^J = \lambda^i a_i$, for any $\mathcal{F}(\hat{x}) \subseteq J \subseteq \mathcal{A}(\hat{x})$ such that $i \in J \setminus \mathcal{F}(\hat{x})$.

We then show that for an arbitrary point, $\tilde{x}$, of a small neighbourhood of $\hat{x}$ such that $a_i^T(\tilde{x} - \hat{x}) \geq 0$, $i \in \mathcal{F}(\hat{x})$, we have:

$$f(\tilde{x}) = f(\hat{x}) + (\tilde{x} - \hat{x})^T c^{\mathcal{A}(\hat{x})} + \sum_{i \in \mathcal{A}(\hat{x}) \setminus \mathcal{F}(\hat{x})} \lambda^i \min(0, a_i^T(\tilde{x} - \hat{x})).$$

Let

$$K \equiv \{i \in \mathcal{A}(\hat{x}) : a_i^T(\tilde{x} - \hat{x}) \geq 0\}.$$

Using the fact that $\mathcal{F}(\hat{x}) \subseteq K$, the rest of the proof then follows exactly as in the proof of proposition 1. □

As in the continuous case, note that the decomposition constructed in the proof is valid for any other point $\tilde{x}$ such that $\sigma(\tilde{x}) = \sigma(\hat{x})$, as long as we assume that artificial ridges were introduced in order to avoid the situation where for a part of a segment we can only possibly drop negatively, while on another part of the same segment we can only possibly drop it positively (specifically, we include as many artificial ridges as necessary so that for any given point $\hat{x} \in I\!R^n$, the vector $\sigma(\hat{x})$ determines $\mathcal{F}^+(\hat{x})$ and $\mathcal{F}^-(\hat{x})$).

4.3 Algorithm

Theorem 5 (Optimality Conditions—Discontinuous Case) *Let $f : I\!R^n \to I\!R$ be a (possibly discontinuous) piecewise linear function with ridges $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$, where $\mathcal{R}$ is a finite index set, and let $x^* \in I\!R^n$ be a non-contact point of f. Assume linear independence of $\{a_i\}_{i \in \mathcal{A}(x^*)}$ and, without loss of generality, $\mathcal{F}(x^*) = \mathcal{F}^-(x^*)$. Let $g_{x^*}, \{\nu_{x^*}^i\}_{i \in \mathcal{A}(x^*) \setminus \mathcal{F}(x^*)}$ be the decomposition of f at x^*. The point x^* is a local minimum of f if and only if there exist scalars u_i^*, $i \in \mathcal{A}(x^*)$ such that*

1. *$g_{x^*} = \sum_{i \in \mathcal{A}(x^*)} u_i^* a_i$ (or, equivalently, the reduced restricted gradient of f at x^* is null, i.e. x^* is a dead point) and*
2. *For each $i \in \mathcal{A}(x^*)$,*

   (i) $u_i^* \leq -\nu_{x^*}^i$, if $i \notin \mathcal{F}(x^*)$

   (ii) $u_i^* \geq 0$.

When x^* is a dead point, condition 2 means that the directional derivative of f in each single-dropping directions which is not a soaring direction, is non-negative.

Proof: The necessity of the first condition holding follows exactly the same argument as in the continuous version of the optimality condition theorem (theorem 2). In order to prove the necessity of condition 2 (i), let $i \in \mathcal{A}(x^*) \setminus \mathcal{F}(x^*)$. Consider moving from a dead point x^* to a displaced point $x^* + \alpha d$, where d satisfies $a_k^T d \geq 0$, $k \in \mathcal{F}(x^*)$. We have, for $\alpha > 0$ small enough:

$$f(x^* + \alpha d) = f(x^*) + \alpha d^T [g_{x^*} + \sum_{k \in \mathcal{A}(x^*) \setminus \mathcal{F}(x^*) : a_k^T d < 0} \nu_{x^*}^k a_k].$$

Using the fact that x^* is a dead point and using $d = d^-$, where d^- drops negatively activity i (we indeed have $a_k^T d^- \geq 0$, for all $k \in \mathcal{F}(x^*)$, as $i \notin \mathcal{F}(x^*)$), we obtain

$$f(x^* + \alpha d^-) = f(x^*) - \alpha(u_i^* + \nu_{x^*}^i)|a_i^T d^-|.$$

If we do not have $u_i^* \leq -\nu_{x^*}^i$, then x^* is not a local minimum of f. We prove the necessity of condition 2 (ii), by using rather a direction, d^+, dropping positively activity i.

Suppose now that both conditions 1 and 2 hold. Consider $d \in I\!R^n$, an arbitrary direction. We want to show that d is not a descent direction. This is clear if $d \in S(x^*)$, thus we can assume that $d \notin S(x^*)$. But this implies $a_i^T d \geq 0$, for all $i \in \mathcal{F}(x^*)$: for suppose that there existed a fault $i \in \mathcal{F}(x^*)$ such that $a_i^T d < 0$. Then, by our convention on the definition of the ridges ($\mathcal{F}(x^*) = \mathcal{F}^-(x^*)$) and the fact that $d \notin S(x^*)$, it means that x^* is a contact point with respect to ridge i (take in the definition (part 2) of a contact point, $\sigma^- = \sigma(x^* + \delta d)$ and $\sigma^+ = \sigma(x^* + \delta \hat{d})$, for $\delta > 0$ small enough, and $\hat{d}$ such that $a_k^T \hat{d} \geq 0, k \in \mathcal{F}(x^*)$ and $a_i^T \hat{d} > 0$). This contradicts the hypothesis that x^* is not a contact point. Hence, using the decomposition of f at x^*, we have, for $\alpha > 0$ small enough:

$$\begin{aligned} f(x^* + \alpha d) &= f(x^*) + \alpha d^T[g_{x^*} + \sum_{i \in \mathcal{A}(x^*) \setminus \mathcal{F}(x^*) : a_i^T d < 0} \nu_{x^*}^i a_i] \\ &= f(x^*) + \alpha \sum_{i \in \mathcal{A}(x^*)} (u_i^* + \lambda_{x^*}^{i,d}) a_i^T d, \end{aligned}$$

since x^* is a dead point, where

$$\lambda_{x^*}^{i,d} = \begin{cases} \nu_{x^*}^i & \text{if } a_i^T d < 0 \text{ and } i \notin \mathcal{F}(x^*), \\ 0 & \text{otherwise.} \end{cases}$$

Thus, by hypothesis and using the fact that $a_i^T d \geq 0$ for all $i \in \mathcal{F}(x^*)$, we have

$$f(x^* + \alpha d) \geq f(x^*).$$

□

From these optimality conditions, we derive an algorithm for minimizing a (possibly discontinuous) piecewise linear function, $f : I\!R^n \to I\!R$ which is very similar to that for the continuous case. Let $\{a_i^T x - b_i\}_{i \in \mathcal{R}}$ be the ridges of f, where $\mathcal{R}$ is a finite index set. We assume again that f is decomposable at each iterate and at each breakpoint encountered in the line search. Moreover, we assume that at every iterate which is a dead point, the gradients of the activities are linearly independent and that all points encountered in the algorithm are non-contact points. In [13], we discuss the degenerate situation and how the algorithm could be modified to take into account contact points.

We assume, without loss of generality, that at the kth iterate, x^k, $\mathcal{F}(x^k) = \mathcal{F}^-(x^k)$.

Discontinuous Piecewise Linear Minimization Algorithm

The only step (from the continuous algorithm) which we need to modify is:

Step 4: For each $i \in \mathcal{A}(x^k)$:
If $u_i < 0$ or $u_i > -\nu_{x^k}^i$, $i \notin \mathcal{F}(x^k)$ (violated optimality condition), then go to step 6.
Otherwise, stop: x^k is a local minimum of f.

The same finite step convergence result is valid, based on theorem 5.

5 Nonlinear Case

Having given the framework of the decomposition developed in the (discontinuous) piecewise linear case, one can consider adapting conventional techniques for nonlinear programming for the general (possibly discontinuous) piecewise differentiable case, as we did above with the projected gradient method for the (possibly discontinuous) piecewise linear case. Indeed, the essence of the concepts introduced in sections 3 and 4 does not rest on the linear nature of the problem.

In this section, we sketch the lines of the extension of our work to the nonlinear case. We define a (possibly discontinuous) *piecewise differentiable function* $f : I\!R^n \longrightarrow I\!R$ to be a function whose derivative is defined everywhere except over a subset of a finite number of sets of the form $\{x \in I\!R^n : r(x) = 0\}$, where r is a differentiable function. In the nonlinear situation, we thus extend the concept of *ridge* to be a specified set $\{x \in I\!R^n : r(x) = 0\}$ containing points where the derivative of f is not defined, where r is a differentiable function (we say also that r is a ridge).

We generalize the definition of decomposition at a point $\hat{x} \in I\!R^n$, so that it expresses the *first-order* behaviour of a piecewise differentiable function in the neighbourhood of $\hat{x}$:

Definition 3 *Let $f : I\!R^n \longrightarrow I\!R$ be a (possibly discontinuous) piecewise* differentiable *function with ridges $\{r_i\}_{i\in\mathcal{R}}$, where $\mathcal{R}$ is a finite index set, and let $\hat{x} \in I\!R^n$. Let $g_{\hat{x}} \in I\!R^n$ and $\psi_{\hat{x}}$ be a function defined on $I\!R^n$ such that we have:*

For all $d \in I\!R^n$ such that $\nabla r_i(\hat{x})^T d \geq 0$, $i \in \mathcal{F}(\hat{x})$, there exists $\delta > 0$ small enough such that:

$$f(\hat{x} + \delta d) = f(\hat{x}) + \delta g_{\hat{x}}^T d + \delta \psi_{\hat{x}}(d) + O(\delta^2),$$

where

$$\psi_{\hat{x}}(d) = \sum_{i\in\mathcal{A}(\hat{x})\setminus\mathcal{F}(\hat{x})} \nu_{\hat{x}}^i \min(0, \nabla r_i(\hat{x})^T d)$$

for some scalars $\{\nu_{\hat{x}}^i\}_{i\in\mathcal{A}(\hat{x})\setminus\mathcal{F}(\hat{x})}$.

We say that $g_{\hat{x}}$, $\{\nu_{\hat{x}}^i\}_{i\in\mathcal{A}(\hat{x})\setminus\mathcal{F}(\hat{x})}$ is a first-order decomposition *of f (into a smooth function and a sum of functions having a single ridge) at $\hat{x}$.*

We would like to impose restrictions on a piecewise differentiable function so that the nonlinearity of its ridges does not prevent the proofs of the decomposition theorem (theorem 4) and of the optimality conditions (which become first-order necessary conditions) to be extended. For instance, in order to be able to adapt standard techniques for nonlinear programming to the piecewise differentiable case, we must restrict our considerations to functions having only a finite number of *pieces* (subdomains over which the function is smooth) in the neighbourhood of a given point (in the piecewise linear case the finiteness of the number of pieces was implied by the finiteness of the number of ridges). A whole class of functions for which the nonlinear analogues of theorems 4 and 5 hold should remain. To give an example, we would expect these results to apply to a function defined over $I\!R^2$ which is quadratic over each of the subdomains delimited by a circle intersecting an ellipse in the plane.

In the remainder of this section, we briefly describe the main issues one should take into account for the extension of our work to the nonlinear case—the general (possibly discontinuous) piecewise differentiable situation.

In sections 3 and 4, the algorithm used descent directions attempting to decrease the smooth part of the function while maintaining the value of its nonsmooth part. A first-order algorithm for the nonlinear case could obtain these two objectives *up to first-order changes*, as in the Conn-Pietrzykowski approach to nonlinear optimization, via a penalty function [14].

In the nonlinear case, the matrix of activities, $A(\hat{x})$, has as its columns the vectors $\{\nabla r_i(\hat{x})\}_{i \in \mathcal{A}(\hat{x})}$. A first remark concerns *near-activities*. In order to avoid zigzagging (see an illustration of this in [8]) when dealing with nonlinear ridges, it is necessary to consider projecting onto the tangent hyperplanes, not only when a ridge, r, is active at the current point, $\hat{x}$, but also when $\|r(\hat{x})\|$ is "small". We consider a ridge, r, *ϵ-active* at $\hat{x}$ whenever $|r(\hat{x})| < \epsilon$, where ϵ is some tolerance (which may be reduced as we approach optimality). This however causes a problem to make active the near-activities as we approach final convergence. The direction of search is thus made up of two components: the projected gradient direction, h^k, is called the *horizontal step* in contrast with the *vertical step*,

$$v^k \equiv -A_\epsilon(x^k)[A_\epsilon(x^k)^T A_\epsilon(x^k)]^{-1}\Phi(x^k + \alpha^k h^k),$$

which attempts to make active the relevant ridges via a linearization ($A_\epsilon(x^k)$ has as its columns the vectors $\{\nabla r_i(x^k)\}_{i \in \mathcal{A}_\epsilon(x^k)}$, where $\mathcal{A}_\epsilon(x^k)$ is the set of the ridges which are near-active at the current iterate, x^k; Φ is the vector of (near-)active ridge values; and α^k is the horizontal step size). (In this context of near-activities, we can see that degeneracy cannot be disregarded for nonlinear problems. In nonlinear l_1 data fitting [2], for example, we may expect many more near-activities than the number of dimensions of the problem if the fit is good, particularly earlier on in many algorithms when ϵ may be relatively large. Hence, degeneracy is very likely to occur in such a problem.) Moreover, in order to establish whether an iterate is degenerate, we use a notion of *linear ϵ-independence*, where ϵ is some small tolerance. For example, we can say that a vector v is linearly ϵ-dependent upon a set of vectors $v_1, \ldots, v_k$ if $\|P(\frac{v}{\|v\|})\| < \epsilon$, where P is the orthogonal projector onto the space orthogonal to the space spanned by $v_1, \ldots, v_k$.

The nonlinear line search algorithm presented in [7] (for the nondifferentiable exact penalty function corresponding to a nonlinear programming problem) can be used here to *estimate* the location of a possible minimum breakpoint along the search direction (in the nonlinear case, the minimum along a search direction need not be a breakpoint). Moreover, one would expect an efficient line search to only find points of sufficient decrease rather than finding the minimum along the line.

The above-mentioned considerations would yield a method converging directly to a local optimum (global convergence), but possessing (in general) only a linear convergence rate (see [14]).

In order to develop a second-order algorithm, assuming now that f is (possibly discontinuous) *piecewise twice-differentiable* (i.e. twice differentiable everywhere except over a finite number of ridges), one must first extend the definition of first-order decomposition to that of *second-order decomposition.*

One could then consider extending the strategies used by Coleman and Conn [7] on the exact penalty function approach to nonlinear programming (although the exact penalty function involves only first-order types of nondifferentiabilities—ridges). The main idea is to attempt to find a direction which minimizes the change in f (up to *second-order terms*) subject to preserving the activities (up to *second-order terms*). Specifically, second-order conditions must be derived [which are the first-order conditions plus a condition on the "definiteness" of the reduced Hessian of the *twice-differentiable part* of f (in the second-order decomposition of f)]. An analogue of the Newton step (or of a modification of Newton's method, see [22]) using a non-orthogonal projection [9] is then taken (or a single-dropping direction is used). An algorithm following these lines would be expected to possess global convergence properties (regardless of starting point) and a fast (2-step superlinear) asymptotic convergence rate as in [6].

6 Conclusion

We introduced in this paper the concepts of the *decomposition* of a function into a smooth part and a nonsmooth part, and the decomposition theorem, which states that a decomposition always exists at non-degenerate points and explicitly gives the decomposition. Optimality conditions and a descent algorithm have then been inferred from the decomposition of the function. Easy generalization of these ideas to the discontinuous situation then followed in section 4 by restricting the decomposition of the function to *non-soaring* directions (at non-contact points). We have hence set a framework for an algorithm dealing directly with discontinuities that could be involved in a non-differentiable optimization problem.

Comparison of our work with recent developments in nondifferentiable optimization, for example Lemaréchal's bundle methods [29] or composite nonsmooth optimization algorithms (see [18]), deserves attention in future work. Particularly, it should be interesting to compare Clarke's more theoretical approach stemming from convex optimization theory and nonsmooth analysis [5], with that introduced in this paper, which is based on methods more oriented to numerical implementations and on practical considerations.

In the second part of this work [13], we discuss the implementation of our algorithm. Namely, we tackle the problem of degeneracy and contact points. We also present encouraging numerical results.

References

[1] R. H. Bartels and A. R. Conn. Linearly constrained discrete l_1 problems. *ACM Transactions on Mathematical Software*, 6(4):594–608, 1980.

[2] R. H. Bartels and A. R. Conn. An approach to nonlinear l_1 data fitting. In J. P. Hennart, editor, *Proceedings of the Third Mexican Workshop on Numerical Analysis*, pages 48–58. Springer Verlag, 1981.

[3] P. H. Calamai and A. Conn. A projected newton method for l_p norm location problems. *Mathematical Programming*, 38(1):75–109, 1987.

[4] C. Cheng and E. Kuh. Module placement based on resistive network optimization. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, 3:218-225, 1984.

[5] F. H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley, 1983.

[6] T. F. Coleman and A. R. Conn. Nonlinear programming via an exact penalty function: Asymptotic analysis. *Mathematical Programming*, 24:123-136, 1982.

[7] T. F. Coleman and A. R. Conn. Nonlinear programming via an exact penalty function: Global analysis. *Mathematical Programming*, 24:137-161, 1982.

[8] A. R. Conn. Constrained optimization using a nondifferentiable penalty function. *SIAM Journal on Numerical Analysis*, 10:760-784, 1973.

[9] A. R. Conn. Projection matrices - a fundamental concept in optimization. In Vogt and Mickle, editors, *7th Annual Conference in Modelling and Simulation, Ed. April 26-27, 1976*, pages 599-605, 1976.

[10] A. R. Conn. Nonlinear programming, exact penalty functions and projection techniques for non-smooth functions. In P. T. Boggs, R. H. Byrd, and R. B. Schnabel, editors, *Numerical Optimization 1984 - Proceedings of the SIAM Conference on Numerical Optimization, Boulder, 1984*, pages 3-25, 1985.

[11] A. R. Conn and G. Cornuéjols. A projection method for the uncapacitated facility location problem. *Mathematical Programming*, 46:273-298, 1990.

[12] A. R. Conn and Y. Li. A structure exploiting algorithm for nonlinear minimax problems. *SIAM Journal on Optimization*, 2(2):242-263, 1992.

[13] A. R. Conn and M. Mongeau. Discontinuous piecewise differentiable optimization II: Degeneracy and applications. Under preparation.

[14] A. R. Conn and T. Pietrzykowski. A penalty function method converging directly to a constrained optimum. *SIAM Journal on Numerical Analysis*, 14(2):348-375, Apr. 1977.

[15] D. De Wolf, O. J. de Bisthoven, and Y. Smeers. The simplex method extended to piecewise-linearly constrained problems I: the method and an implementation. Technical report, Center for Operations Research and Econometrics, Louvain-La-Neuve, Belgium, 1991.

[16] D. De Wolf, O. J. de Bisthoven, and Y. Smeers. The simplex method extended to piecewise-linearly constrained problems II: an application to the gas transmission problem. Technical report, Center for Operations Research and Econometrics, Louvain-La-Neuve, Belgium, 1991.

[17] S. S. Erenguc and H. P. Benson. The interactive fixed charge linear programming problem. *Naval Research Logistics Quarterly*, 33:157-177, 1986.

[18] R. Fletcher. *Practical Methods of Optimization.* Wiley-Interscience, second edition, 1987.

[19] R. Fourer. A simplex algorithm for piecewise-linear programming I: Derivation and proof. *Mathematical Programming*, 33:204–233, 1985.

[20] R. Fourer. A simplex algorithm for piecewise-linear programming III: Computational analysis and applications. Technical Report 86-03, Dept. of Industrial Engineering and Management Sciences, Northwestern University, Evanston IL, 1986. (Revised 1988, 1989).

[21] R. Fourer. A simplex algorithm for piecewise-linear programming II: Finiteness, feasibility and degeneracy. *Mathematical Programming*, 41:281–315, 1988.

[22] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization.* Academic Press, 1981.

[23] J. Hald and K. Madsen. Combined LP and Quasi-Newton methods for minimax optimization. *Math. Prog.*, 20:49-62, 1981.

[24] S. P. Han. Variable metric methods for minimizing a class of nondifferentiable functions. *Mathematical Programming*, 20:1–13, 1981.

[25] S. Hiraki. A simplex procedure for a fixed charge problem. *Journal of the Operations Research Society of Japan*, 23(3):243-266, 1980.

[26] I. I. Imo and D. J. Leech. Discontinuous optimization in batch production using sumt. *International Journal of Production Research*, 22(2):313-321, 1984.

[27] E. M. Klein and S. H. Sim. Discharge allocation for hydro-electric generating stations. *European Journal of Operational Research*, 1992. Submitted for possible publication.

[28] D. E. Knuth. *The Art of Computer Programming, Vol. 3, Sorting and Searching.* Addison-Wesley, 1975.

[29] C. Lemaréchal. Bundle methods in nonsmooth optimization. In C. Lemaréchal and R. Mifflin, editors, *Proceedings of the IIASA Workshop, Nonsmooth Optimization, March 28-April 8, 1977*, volume 3, pages 79–102. Pergamon Press, 1978.

[30] M. Mongeau. *Discontinuous Piecewise Linear Optimization.* PhD thesis, Dept. of Combinatorics & Optimization, University of Waterloo, Ontario, Canada, 1991.

[31] B. Montreuil, H. D. Ratliff, and M. Goetschalckx. Matching based interactive facility layout. *AIIE Transactions*, 19(3):271–279, 1987.

[32] W. Murray and M. Overton. A projected Lagrangian algorithm for nonlinear l_1 optimization. *SIAM J. Sci. Statist. Comput.*, 2:207–224, 1981.

[33] M. R. Osborne and G. A. Watson. An algorithm for minimax approximation in the nonlinear case. *Computing J.*, 12:63–68, 1968.

[34] U. S. Palekar, M. H. Karwan, and S. Zionts. A branch-and-bound method for the fixed charge transportation problem. *Management Science*, 36(9):1092–1105, 1990.

[35] A. Tishler and I. Zang. A switching regression method using inequality conditions. *Journal of Econometrics*, 11:259–274, 1979.

[36] G. Vijayan and R.-S. Tsay. A new method for floorplanning using topological constraint reduction. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, 1991. To appear in October 1991.

[37] R. S. Womersley and R. Fletcher. An algorithm for composite nonsmooth optimization problems. *J. Opt. Theory and Applis.*, 48:493–523, 1986.

[38] I. Zang. Discontinuous optimization by smoothing. *Mathematics of Operations Research*, 6(1):140–152, 1981.

[39] J. Zowe. Nondifferentiable optimization. In K. Schittkowski, editor, *Computational Mathematical Programming*, pages 323–356. NATO Advanced Science Institute Series F: 15, Bad Windsheim, July 1984, Springer-Verlag, 1985.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique

ISSN 0249-6399