On Lipschitz optimization based on gray-box piecewise linearization

Andreas Griewank[1], Andrea Walther[2],

Sabrina Fiege[2], and Torsten Bosse[1]

February 10, 2015

Abstract

We address the problem of minimizing objectives from the class of piecewise differentiable functions whose nonsmoothness can be encapsulated in the absolute value function. They possess local piecewise linear approximations with a discrepancy that can be bounded by a quadratic proximal term. This overestimating local model is continuous but generally nonconvex. It can be generated in its *abs-normal form* by a minor extension of standard algorithmic differentiation tools. Here we demonstrate how the local model can be minimized by a bundle type method, which benefits from the availability of additional *gray-box information* via the abs-normal form. In the convex case our algorithm realizes the consistent steepest descent trajectory for which finite convergence was established in [17], specifically covering counter examples where steepest descent with exact line-search famously fails. The analysis of the abs-normal representation and the design of the optimization algorithm is geared towards the general case, whereas the convergence proof so far only covers the convex case.

Keywords

Bundle methods, Piecewise linearity, Algorithmic differentiation, Abs-normal form, Nonsmooth Optimization

[1]Department of Mathematics, Humboldt University of Berlin, Berlin
[2]Department of Mathematics, University of Paderborn, Paderborn

1 Motivation, background, and notation

In scientific computing problem specific functions are almost always evaluated by computer programs composed of elementary functions. If all of them are locally smooth then proper derivatives can be obtained by algorithmic differentiation [13]. However, in many application models nonsmoothness arises through cut-offs at certain minimal or maximal levels of intermediate quantities. An example are slope limiters in the discretization of hyperbolic differential equations. Similarly, one may take maxima or minima of various quantities, e.g., to bound the usage of utilities in economics or to avoid self-penetration of geometrical objects. Finally, the reformulation of constrained problems into an unconstrained form by adding ℓ_1- or ℓ_∞-penalty terms of the constraint violations to the original objective yield nonsmooth unconstrained optimization problems of the kind considered here.

Despite the importance for a wide range of fields, there is still a scarcity of practical methods for the unconstrained minimization of Lipschitzian functions $f : \mathbb{R}^n \mapsto \mathbb{R}$, even in the convex case. Therefore, our long-term goal is the development and analysis of an algorithm for the minimization of general piecewise smooth functions that are continuous but not necessarily convex. In the present paper, we pursue a successive piecewise linearization approach and concentrate on the minimization of piecewise linear (PL) problems by a bundle-type method. In a forthcoming paper [14], we will complete the convergence theory and algorithmic design for the general, nonconvex and nonlinear case.

Lessons from a paradigmatic example

Since the objective function is not differentiable, one possible solution approach is based on derivative-free methods as proposed for example in [16]. However, as we want to consider piecewise smooth function, the exploitation of derivative information promises benefits for the optimization process.

Several texts on nonsmooth optimization (see, e.g., [1] and [3]) highlight examples of convex unconstrained minimization problems, where the steepest descent method with exact line-searches exhibits zigzagging convergence to a nonstationary point. Since in the smooth case this variant of steepest descent is considered quite reliable (if a bit slow) that observation seems rather discouraging. For this reason, variants of the steepest descent method were derived (see, e.g., [5]) to exploit a special form of nonsmoothness.

In view of its sublinear rate of convergence the alternative (see, e.g., [26, Chap. 2]) of proceeding along the negatives of arbitrary subgradients using a sequence of merely square summable step lengths does not really seem enticing

either. As alternative, one could adapt quasi-Newton methods for the nonsmooth case as proposed in [20]. Also a more reasonable rate of convergence can be expected from bundle-methods (see, e.g., [6, 19, 21]), but their performance is somewhat erratic. We try to overcome this by providing our bundle variant with additional information about the objective that is readily available through an extension of automatic, or algorithmic differentiation. In this way we can realize the method originally investigated by Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal in the first volume of the seminal book on *Convex Analysis and Optimization Algorithms* [17] by exploiting directional derivatives. Throughout this paper we stay in a finite dimensional setting, generalizations to Banach spaces were for example considered in [10, 22].

Hiriart-Urruty and Lemaréchal highlighted the piecewise linear, convex example function $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x_1, x_2) = \max\{-100, 3x_1 - 2x_2, 3x_1 + 2x_2, 2x_1 - 5x_2, 2x_1 + 5x_2\}. \qquad (1)$$

Curiously, most authors only cite the bad news of nonconvergence for a standard steepest descent variant that yields a sequence of iterates converging to the point $\bar{x} = (0, 0)$, which is not even stationary. The resulting sequence is depicted in Fig. 1, where the gray shaded area marks the set of optimal points. The same zigzagging-effect occurs on the piecewise quadratic Wolfe example [28].

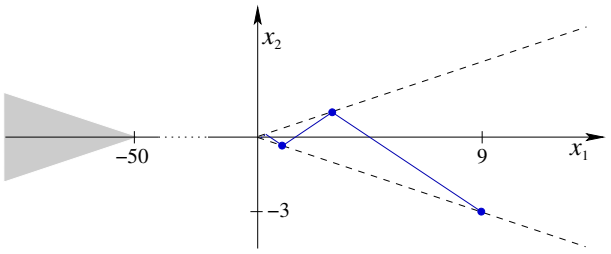


Figure 1: Iterates of the steepest descent method for example (1)

Content and organization of the paper

It seems to be consistently overlooked that in [17] one also finds the good news, namely that in the convex case the continuous steepest descent trajectory defined by the differential inclusion (see, e.g., [2])

$$-\dot{x}(t) \equiv -\frac{d}{dt}x(t) \in \partial f(x(t)) \qquad (2)$$

is unique and does converge to a stationary point and thus a minimizer, provided f attains its infimum as a minimum. Whereas this result was apparently considered merely theoretical, it is the basis for our implementable bundle-like optimization algorithm. For this new approach, we show the convergence in finitely many steps for a piecewise linear, convex objective function with a proximal term. In these cases, our *safe steepest descent* algorithm exactly generates the unique solution trajectory of (2) discussed in [17].

As shown already in [11] the algorithm proposed here can serve as an inner loop in combination with quadratic overestimation of a successive piecewise linearization method to minimize also piecewise smooth objective functions using a proximal term. For this reason, we present in addition to the numerical results on piecewise linear objective functions also the minimization of a Lipschitzian piecewise smooth function given by Nesterov's nonsmooth version of Rosenbrock [15].

The paper is organized as follows. In Section 2 we discuss the crucial issue of what information about a general nonsmooth objective function $f(x)$ can be reasonably expected to be provided by the user. Here, we recommend a shift of paradigm from the usual black-box oracle to a gray-box interface based on information that can be provived for example by an appropriated extended algorithmic differentiation tool. In Section 3, we analyze stationarity and first order minimality of a locally Lipschitzian f at a given point x and discuss algorithms to decide whether these properties are attained at that point. The stationarity test uses a bundle $G \subset \partial f(x)$ and yields a descent direction if the test fails. From Section 4 onward, we concentrate on PL objective functions. For this class of objectives, Section 4 introcudes a representation in the so-called abs-normal form and the corresponding polyhedral decomposition. In Section 5, we propose an optimization algorithm, based on the descent direction derived in Section 3, for PL functions with a proximal term and show its convergence in a finite number of iterations for the convex case. First numerical results are presented in Section 6. Finally, in Section 7 we give a conclusion and an outlook.

Notation and theoretical background

Throughout this paper the multi-function $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ denotes the subdifferential for convex f and the generalized gradient in the sense of Clarke for locally Lipschitzian real-valued functions $f : \mathbb{R}^n \mapsto \mathbb{R}$, see [4, Chap. II]. The fact that ∂f is closed, convex, and outer semi-continuous ensures by Theorem 1.4 in [2, Chap. II] that at least one absolutely continuous solution $x(t)$ of the autonomous differential inclusion (2) exists for each initial condition $x(0) = x_0 \in \mathbb{R}^n$. Moreover, as stated in Theorem 3.4.1 in [17, Chap. VIII], the monotonicity of ∂f in the

convex case ensures not only that $x(t)$ is unique but also that its right derivatives satisfies for all $t \in \mathbb{R}$

$$D_+x(t) = \lim_{\Delta t \searrow 0} \tfrac{1}{\Delta t}[x(t+\Delta t) - x(t)] = d(x(t)),$$

where the vector function $d : \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$d(x) \equiv \text{short}\,(0, -\partial f(x)) \equiv \text{argmin}\,\{\|d\| \mid -d \in \partial f(x)\}\,. \tag{3}$$

Here and throughout the paper $\|\cdot\|$ denotes the Euclidean norm, whose strict convexity ensures that for any vector $h \in \mathbb{R}^n$ and closed subset $G \subset \mathbb{R}^n$ there is a unique singleton

$$\text{short}(h, G) \equiv \text{argmin}\left\{\|d\| \,\middle|\, d = \sum_{j=1}^{m} \lambda_j g_j - h, g_j \in G, \lambda_j \geq 0, \sum_{j=1}^{m} \lambda_j = 1\right\}. \tag{4}$$

The fundamental importance of the descent direction

$$\text{short}\,(0, -\partial f(x)) = -\text{short}\,(0, \partial f(x)) = -P_0(\partial f(x))$$

with P_z the Euclidean projection onto the point $z \in \mathbb{R}^n$ was already discussed by Lemaréchal in his contribution to the classical collection [18].

The role of $d(x)$ is illustrated in Fig. 2 when $\partial f(x)$ contains the three vectors $\{g_1, g_2, g_3\}$ and the convex hull $\text{conv}(-\partial f(x))$ is the gray shaded area. Then, $d(x)$ is given as the projection of $0 \in \mathbb{R}^n$ onto this convex set.

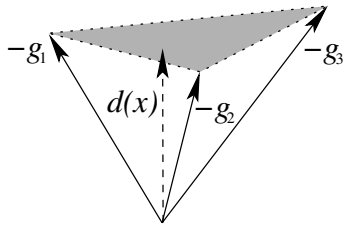


Figure 2: Direction $d(x) = \text{short}(0, -\partial f(x))$

In general, it is even in the convex case not clear how the steepest descent trajectory $x(t)$ can be traced or at least approximated algorithmically. Furthermore, if a minimizer $x_* = x(t_*)$ is indeed reached at some time $t_* < \infty$ the trajectory may have an infinite number of direction changes at times t_j with a limit point $\lim_{j\to\infty} t_j < t_*$. Then a practical algorithm can never reach the minimizer, just

like the zigzagging steepest descent sequence on the counter example mentioned above. Such *Zeno* behavior has received considerable attention in the literature on switching systems, e.g. [25]. On piecewise smooth problems a Zeno effect can be definitely ruled out, at least if we also have convexity. More specifically, Algorithm 3.4.6 in [17] exactly traces the true descent trajectory $x(t)$ as solution of (2). Furthermore, it is shown in [17] that their Algorithm 3.4.6 converges in finitely many steps for any convex PL function that is bounded below. Curiously, this method is rarely mentioned in the literature and apparently considered not implementable because the knowledge of the full set $\partial f(x)$ seems to be required to yield the guaranteed descent direction $d(x)$ defined in (3).

We will show in the present paper that this is not the case. That is, we base our optimization algorithm for convex PL functions on a bundle that may be only a proper subset of $\partial f(x)$. This is the major difference and the important extension of the already known Algorithm 3.4.6 in [17] which makes the optimization algorithm also implementable. Furthermore, one has to note that the algorithm proposed here differs from the simplex method in that more than one constraint can be released in any iteration. Hence, it is not an adaption of the simplex method for PL function as described for example in [8].

2 From black-box oracle to gray-box interface

Much of the algorithmic design and theoretical analysis in nonsmooth optimization is predicated on the *black-box* assumption that all the user can provide about the function to be optimized is an oracle yielding a scalar-vector pair of values

$$f(x) \in \mathbb{R} \quad \text{and} \quad g \in \partial f(x) \subset \mathbb{R}^n \quad \text{at any} \quad x \in \mathbb{R}^n. \tag{5}$$

We deem this to be a rather incongruous scenario for the following reasons:

1. By Rademacher's theorem f possesses almost everywhere a proper gradient, so coding up anything else seems for the most part a wasted effort required of the user. After all, very few iterates of an iterative algorithm are likely to belong to the exceptional set of nondifferentiability often denoted by Θ. In fact in [20] that is simply assumed never to happen.

2. The information provided by the oracle (5) is strictly local and does not yield indications of any nearby nonsmoothness. In particular, there may be no hint of a local minimizer being in the immediate vicinity, which would be required for effective stopping criteria.

3. Contrary to the impression created in part of the nonsmooth literature, computing at any exceptional point $x \in \Theta$ just one vector $g \in \mathbb{R}^n$ that is guaranteed to be a generalized gradient, i.e., $g \in \partial f(x)$, may be difficult. The reason is that simple chain ruling generally does not work, as can be seen easily for example on the expression $f(x) = |x + |x|| - |x|$ at $x = 0$.

4. If one goes through the trouble of providing mechanisms for properly evaluating generalized gradients one then obtains in fact much more information that can be used to reduce much of the uncertainty and heuristics in bundle method design.

At least in theory some of the shortcomings mentioned above can be overcome by considering ε-gradients $\partial_\varepsilon f(x) \supset \partial f(x)$ as defined in [26, Chap. 2]. Naturally, their practical approximation is anything but trivial and of course a fortuitous choice of the tolerance parameter $\varepsilon > 0$ is crucial for algorithmic progress.

Throughout, we will make the entirely realistic assumption that the underlying function $f(x)$ is evaluated by a sequence of elementary operations that are all either Lipschitz continuously differentiable in the domain $D \subset \mathbb{R}^n$ of interest or can be expressed in terms of the absolute value function $v = |u|$. Using reformulations like

$$\max(x, y) = \tfrac{1}{2}(x + y + |x - y|), \tag{6}$$

the usual sources of nonsmoothness, like minima, maxima and complementarity conditions can be rewritten in terms of the absolute value function. Consequently, the considered objective function $f(x)$ is piecewise smooth in the sense of Scholtes [24, Chap. 4]. Assuming that s absolute value functions are evaluated during the calculation of $f(x)$ and collecting their arguments in one vector $z = z(x) : \mathbb{R}^n \to \mathbb{R}^s$, one can define the signature vector

$$\sigma \equiv \sigma(x) \equiv \operatorname{sign}(z(x)) \in \{-1, 0, 1\}^s. \tag{7}$$

Then, the objective function f may be written in the form

$$f(x) \in \{f_\sigma(x) : \sigma \in \mathcal{E} \subset \{-1, 0, 1\}^s\} \quad \text{at} \quad x \in \mathbb{R}^n,$$

where the selection functions f_σ are continuously differentiable on neighborhoods of points where they are active, i.e., coincide with f. We will assume that all f_σ with $\sigma \in \mathcal{E}$ are essential in that their coincidence sets $\{f(x) = f_\sigma(x)\}$ are the closures of their interiors.

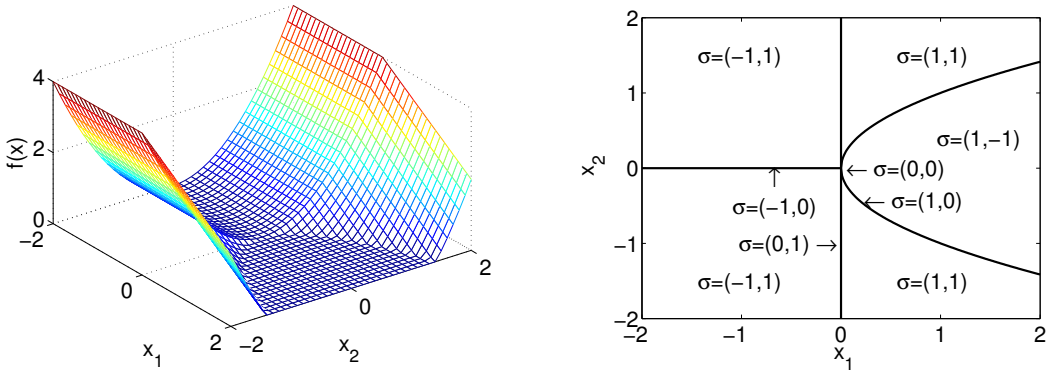


Figure 3: Example function (8) and its signature vector $\sigma(x)$

Example 2.1. *Using the reformulation as stated in* (6)*, the calculation of the function*

$$f : \mathbb{R}^2 \to \mathbb{R}, \qquad f(x_1, x_2) = (x_2^2 - (x_1)_+)_+ \quad \textit{with} \quad y_+ \equiv \max(0, y), \tag{8}$$

at any given point $x = (x_1, x_2)$ *involves two evaluations of the absolute value functions. The function itself and the values of the signature vectors* $\sigma(x)$ *are illustrated by Fig. 3.*

It follows immediately that the generalized gradient is given by

$$\partial f(x) \equiv \mathrm{conv}(\partial^L f(x)) \quad \text{with} \quad \partial^L f(x) \equiv \{\nabla f_\sigma(x) : f_\sigma(x) = f(x)\}.$$

We will call the elements of $\partial^L f(x)$ the limiting gradients of f at x. Finally, as shown in [11, Sect. 3.1, Sect. 3.2], one can obtain constructively a piecewise linear approximation $\Delta f(x; \Delta x)$, which is generally nonhomogeneous and satisfies

$$\Delta f(x; d) = f(x + d) - f(x) + \mathcal{O}(\|d\|^2). \tag{9}$$

In other words, we have a generalized Taylor expansion of first order at x. For particular classes of problems such piecewise linearizations have been considered quite frequently in the literature. A very important aspect of this approximation is that it varies continuously with respect to the base point x in that

$$[\Delta f(\tilde{x}; d) - \Delta f(x; d)]/(1 + \|d\|) = \mathcal{O}(\|\tilde{x} - x\|),$$

cf. Proposition 1 of [11].

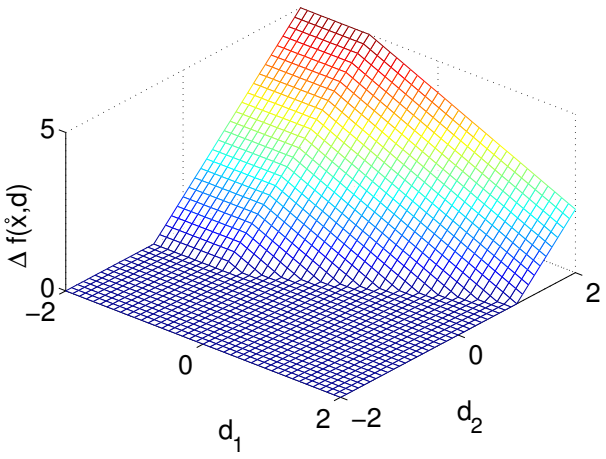


Figure 4: Piecewise linearization (10) of (8) at $\mathring{x} = (1, 1)$

Example 2.2. *Using the approach proposed in [11, Sect. 3.1], one obtains for the nonlinear function f defined in* (8)*, a given base point $\mathring{x}$, and the argument d the piecewise linearization*

$$\Delta f(\mathring{x}; d) = \frac{1}{2}\left(\mathring{x}_2^2 + 2\mathring{x}_2 d_2 - \frac{1}{2}\left(\mathring{x}_1 + d_1 + |z_1|\right) + \frac{1}{2}|z_2|\right),$$
$$z_1 = \mathring{x}_1 + d_1, \qquad z_2 = \mathring{x}_2^2 + 2\mathring{x}_2 d_2 - \frac{1}{2}\left(\mathring{x}_1 + d_1 + |z_1|\right),$$

with the intermediate values z_i as arguments of the absolute value function as introduced above.

Hence, at the origin $\mathring{x} = (\mathring{x}_1, \mathring{x}_2) = 0 \in \mathbb{R}^2$, a piecewise linear approximation of f is given by $\Delta f(0; d) \equiv 0$. This fits nicely to the corresponding illustration given in Fig. 3. When one derives the piecewise linearization at $\mathring{x} = (\mathring{x}_1, \mathring{x}_2) = (1, 1) \in \mathbb{R}^2$ one obtaines for $d = (d_1, d_2)$

$$\Delta f(\mathring{x}; d) = \frac{1}{4} - \frac{d_1}{4} + d_2 - \frac{|1 + d_1|}{4} + \frac{1}{2}\left|\frac{1}{2} - \frac{d_1}{2} + 2d_2 - \frac{|1 + d_1|}{2}\right|. \tag{10}$$

This local piecewise linear modell is illustrated in Fig. 4.

Under our assumptions of piecewise smoothness the directional derivative

$$f'(x; d) = \lim_{\tau \searrow 0} \tfrac{1}{\tau}[f(x + \tau d) - f(x)] \tag{11}$$

is well defined for all pairs $x, d \in \mathbb{R}^n$. Moreover it is piecewise linear with

$$f'(x; \tau\, d) = \Delta f(x; \tau\, d) \quad \text{for} \quad \tau \gtrsim 0, \tag{12}$$

see [11, Sect. 3.2]. In other words, $f'(x; d)$ is the homogeneous part of the piecewise linear approximation $\Delta f(x; d)$.

As detailed in [11] and [12] on our gray-box scenario, the following information can be readily computed at any pair $x, d \in \mathbb{R}^n$ with $d \neq 0$ with an appropriate extended algorithmic differentiation:

1. A *directionally active gradient* $g \equiv g(x; d) \in \partial^L f(x)$ such that $f'(x; d) = g^\top d$ and $g(x; d)$ equals the gradient $\nabla f_\sigma(x)$ of a locally differentiable selection function f_σ that coincides with f on a set, whose tangent cone at x contains d and has a nonempty interior.

2. The value $\Delta f(x, d)$ and a maximal *critical multiplier* $\hat{\tau} \in (0, \infty]$ such that $\Delta f(x, \tau\, d) = \tau\, g^\top d$ for $0 \leq \tau < \hat{\tau}$.

3. Directionally active gradients and critical multipliers on the shifted piecewise linear approximation $\Delta_x f(\tilde{x}; d) \equiv \Delta f(x, \tilde{x} - x + d)$ with $\tilde{x}$ fixed. We will denote them by $g_x(\tilde{x}; d)$ and $\hat{\tau}_x(\tilde{x}, d)$.

Using the reverse mode of algorithmic differentiation [13, Sect. 3.2] one obtains directionally active gradients normally at roughly the same cost as evaluating $f(x)$ itself. However, the cost ratio may grow up to $\mathcal{O}(n)$ in very degenerate circumstances. The cost of computing the critical multiplier $\hat{\tau}$ is always of the same order as that of evaluating f itself. General purpose drivers to compute the directional active gradients and the critical multiplier will be contained in the next version of the AD-tool ADOL-C [27].

Our first objection to the usual black-box paradigm, namely that f is almost everywhere differentiable so that $g(x; d)$ is simply the conventional gradient $\nabla f(x)$ still applies. However, when the critical multiplier $\hat{\tau} = \hat{\tau}(x, d)$ is finite the directionally active gradient $g_x(x + \hat{\tau} d; \tilde{d})$ is likely to differ from $\nabla f(x; d)$ for most $\tilde{d}$ including $\tilde{d} = d$. In this way one obtains approximate gradients that apply in the vicinity of the base points x and may in fact be ε-gradients.

3 Stationarity and first order optimality

In the convex case, the first order minimality condition is satisfied at a given point x if and only if the point is *stationary* in that

$$0 \in \partial f(x) \quad \Longleftrightarrow \quad 0 = d(x) = -\text{short}\,(0, \partial f(x))$$

with short(.) as defined in (4). At all nonstationary points we then have the unique direction of steepest descent

$$d(x)/\|d(x)\| \;=\; \operatorname{argmin}\{f'(x;e) : \|e\| \leq 1\}.$$

From these observations, we obtain the following simple algorithm to compute the direction of the next step in our gray-box scenario.

Algorithm 3.1 (Step Computation I)**.**

ComputeStep(x**,** G**)** *// Precondition:* $x \in \mathbb{R}^n$, $\emptyset \neq G \subset \partial^L f(x)$
repeat
  { $d = -short(0, G)$
    $g = g(x; d)$
    $G = G \cup \{g\}$
  }
until $g^\top d \leq -\|d\|^2$
eliminate all $\tilde{g} \in G$ *with* $\tilde{g}^\top d \neq g^\top d$
return d, G

As we see, Algo. 3.1 returns for a given point x a direction d and a possibly modified $G \subset \partial^L f(x)$. The old limiting gradients $\tilde{g}$ that do not have the same inner product with d as the final g must be eliminated because they cannot be active at the points $x + \tau\, d$ even for small $\tau > 0$. In the convex case there can be no $\tilde{g} \in \partial^L f(x)$ with $\tilde{g}^\top d > g^\top d$ due to the fact that the last computed g is directionally active, i.e., $g^\top d = -\|d\|^2 = f'(x; d)$. Irrespective of convexity properties we obtain the following result.

Proposition 3.2 (Safe Descent)**.** *Algorithm 3.1 terminates after finitely many iterations. On return* $d = 0$ *implies that* f *is stationary at the input point* x*, i.e.,* $0 \in \partial f(x)$*. Otherwise, the return vector* d *is a direction of safe descent in that*

$$f'(x; d) \;\leq\; -\|d\|^2 \;<\; 0. \tag{13}$$

Moreover, when f *is convex, we have minimality of* x *if and only if* $d = 0$ *and otherwise* $d = d(x)$ *is the up to scaling unique direction of steepest descent at* x*.*

Proof. Since $G \subset \partial^L f(x)$ is monotonically enlarged and $\partial^L f(x)$ contains only a finite number of elements, the norm $\|d\|$ is monotonically decreasing and must reach a minimum after finitely many iterations. If $d = 0$ we clearly must have stationarity.

When on exit $d \neq 0$ holds, this vector represents a descent direction since for $g = g(x,d) \in G$ by definition of d and elementary convex geometry one obtains

$$f'(x;d) = g^\top d \leq -\|d\|^2 < 0.$$

In the convex case the stationarity $d = 0$ is equivalent to first order minimality. If $d \neq 0$ the convexity of f ensures that d is the direction of steepest descent. $\square$

We will refer to the property (13) as generalized steepest descent. If at a resulting sequence of iterates the function values are bounded below and the step multipliers do not converge to zero we can then conclude that there must be a stationary cluster point.

The number of iterations required by Algorithm 3.1 is bounded by 3^s, i.e., the maximal number of selection functions, which may theoretically all be active at x. Furthermore, it is important to note that in the convex case using Algorithm 3.1 the decision whether x is a stationary point of f can be made without the guarantee that the full set $\partial^L f(x)$ has been computed. In the nonconvex case one obtains either a direction of descent or the information that x is stationary.

While in the convex case $d = 0$ ensures optimality of x, we know in the nonconvex case only that $\mathrm{conv(G)} \subset \partial f(x)$ and thus for arbitrary $e \in \mathbb{R}^n$

$$f'(x;e) \;\leq\; \max\{g^\top e : g \in \partial f(x)\} \;\geq\; 0$$

so that the existence of a descent direction cannot be excluded. For example, the simple function $f(x) = -|x|$ has $\partial f(0) = [-1,1]$ and thus $\mathrm{short}(0, -\partial f(0)) = 0$ but there are two direction of steepest descent namely -1 and 1. Hence, in the nonconvex case $d(x) = \mathrm{short}\,(0, -\partial f(x))$ should be more appropriately called the direction of *safe descent* as introduced in Prop. 3.2 since one obtains easily

$$d(x)/\,\|d(x)\| \in \operatorname*{argmin}_{\|e\|\leq 1} \max_{g\in\partial f(x)} g^\top e.$$

As can be seen already for the simple example $f(x) = -|x|$, stationarity is a much weaker property than first order minimality. Correspondingly, even while Algorithm 3.1 may of course take some time, testing first order optimality in the nonconvex case is much more difficult. In effect we must globally minimize for fixed x with respect to a unit vector e the function $f'(x;e)$, which may be a completely general homogeneous linear function in the n components of e.

Partitioning $e = (e_1, e_{-1})$ with $e_1 \in \mathbb{R}$ the first component of e we could equivalently globally minimize with respect to $e_{-1} \in \mathbb{R}^{n-1}$ the three PL functions

$$f'(x;(-1,e_{-1})), \quad f'(x;(1,e_{-1})), \quad \text{and} \quad f'(x;(0,e_{-1})).$$

again subject to constraints on the norm of e_{-1} and the modulus of e_1. While $f'(x;(0,e_{-1}))$ is again homogeneous, the other two are not, so that we would have to globally minimize two inhomogeneous and one homogeneous PL functions in $n-1$ variables. These could be treated by iterative methods based on local descent requiring first order optimality tests in $n-2$ variables and so on. It seems clear that such a recursion in the dimension would lead to an enormous combinatorial effort, which could only be worthwhile in rare situations.

When x is in fact first order minimal looking for a descent direction in the way sketched above would involve 3^n recursive calls. If x is not first order minimal one might strike it lucky and find a descent after just n recursive calls if one uses a depth first strategy to traverse the ternary calling tree. The exponential complexity is no surprise since 3 SAT [9] can be posed as the decision problem whether the global minimum of a corresponding PL function is zero. For this reason, we will restrict our ambition to just locating stationary points in the remainder of this paper.

4 The PL objective in abs-normal form

From now on, we will consider only PL functions $f : \mathbb{R}^n \to \mathbb{R}$. Therefore, we will use x and no longer d as argument of the PL function to simplify notation. Furthermore, since we wish to minimize let us assume for simplicity that $f(0) = 0$. As shown in [12, Sect. 2] any such PL scalar function $y = f(x)$ can be expressed in terms of a so-called switching vector $z \in \mathbb{R}^s$ in the *abs-normal* form

$$\begin{bmatrix} z \\ y \end{bmatrix} = \begin{bmatrix} c \\ 0 \end{bmatrix} + \begin{bmatrix} Z & L \\ a^\top & b^\top \end{bmatrix} \begin{bmatrix} x \\ |z| \end{bmatrix}, \tag{14}$$

where

$$c \in \mathbb{R}^s,\ Z \in \mathbb{R}^{s\times n},\ L \in \mathbb{R}^{s\times s},\ a \in \mathbb{R}^n,\ b \in \mathbb{R}^s.$$

The matrix L is strictly lower triangular, i.e., each z_i is an affine function of absolute values $|z_j|$ with $j < i$ and the independents x_k for $1 \le k \le n$. Thus we have a piecewise linear vector function $z = z(x) : \mathbb{R}^n \to \mathbb{R}^s$. It is important to note once more that an extended version of algorithmic differentiation as implemented already in ADOL-C can be used to compute the abs-normal form. Therefore, one only has to provide a corresponding computer programm evaluating f to obtain also an abs-normal form of the objective function.

Example 4.1. *Considering again the piecewise linearization* $\Delta f(\mathring{x}; x)$ *of the objective function* (8) *at* $\mathring{x} = (\mathring{x}_1, \mathring{x}_2) = (1, 1) \in \mathbb{R}^2$ *as stated in* (10)*. The corresponding abs-normal form is given by*

$$\begin{bmatrix} z_1 \\ z_2 \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \\ 1/4 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1/2 & 2 & -1/2 & 0 \\ -1/4 & 1 & -1/4 & 1/2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ |z_1| \\ |z_2| \end{bmatrix}. \tag{15}$$

One can check very easily that the sets

$$P_\sigma \equiv \{x \in \mathbb{R}^n : \sigma(x) = \sigma\} \tag{16}$$

are relatively open and convex polyhedra in $\mathbb{R}^n$. Being inverse images they are mutually disjoint and their union is the whole of $\mathbb{R}^n$. We may define the property of P_σ being relatively open as not having a proper convex subset whose closure contains P_σ. In that sense, single points are also relatively open. Of course minima of piecewise linear functions may be attained not only at single points, but the proximal term considered later may generate isolated local minima within the relative interior of higher dimensional polyhedra.

By continuity it follows that P_σ must be open (but possibly empty) if σ is *definite* in that all its components are nonzero. Whenever σ contains zero entries it is called *critical*. In degenerate situations there may be some critical σ that are nevertheless *open* in that P_σ is open. The set of all polyhedra P_σ form a directed acyclical graph, which is called a skeleton by Scholtes, see [24, Chap. 2].

Now let us freeze any $\sigma \in \{-1, 0, 1\}^s$ and substitute $|z| \equiv \Sigma z$ using the signature matrix $\Sigma = \mathrm{diag}(\sigma)$ defined by

$$\Sigma \equiv \Sigma(x) \equiv \mathrm{diag}(\sigma) \ \in \{-1, 0, 1\}^{s \times s}$$

for the signature vector $\sigma(x)$ (see also [11, Sect. 6.1]). Then the first equation in (14) yields

$$(I - L\Sigma)z = c + Zx \quad \text{and} \quad z \ = \ (I - L\Sigma)^{-1}(c + Zx). \tag{17}$$

Notice that due to the strict triangularity of $L\,\Sigma$ the inverse of $(I - L\Sigma)$ is well defined and polynomial in the entries of L.

Substituting this expression into the last equation of (14) we obtain the selection function

$$f_\sigma(x) \equiv \gamma_\sigma + g_\sigma^\top x \tag{18}$$

with

$$\gamma_\sigma = b^\top \Sigma (I - L\Sigma)^{-1} c \quad \text{and} \quad g_\sigma^\top = a^\top + b^\top \Sigma (I - L\Sigma)^{-1} Z. \tag{19}$$

We certainly have by definition of $\sigma = \sigma(x)$

$$\overline{P}_\sigma \ \subset \ \{x \in \mathbb{R}^n : f(x) = f_\sigma(x)\}$$

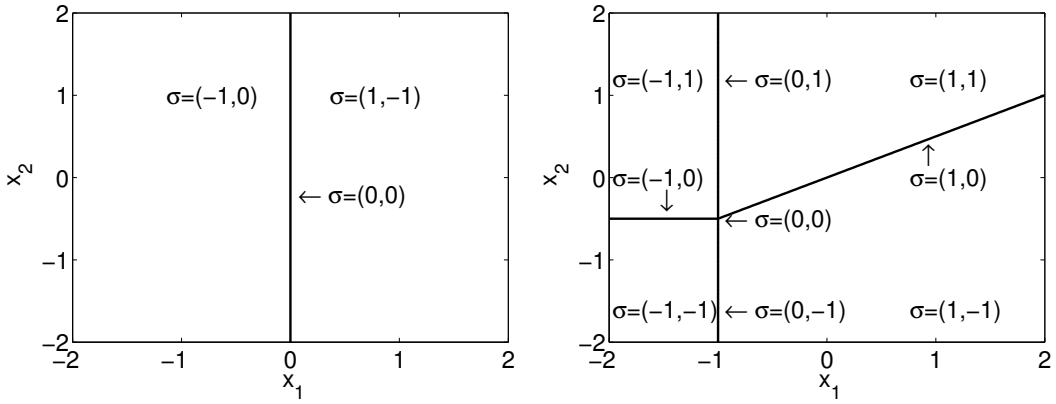
where equality must hold in the convex case. In the nonconvex case, f_σ may coincidentally be active, i.e., coincide with f at points in other polyhedra $P_{\tilde{\sigma}}$. In fact the coincidence sets may be the union of many polyhedral components but given the abs-normal form there is no need to deal with any of its arguments outside $\overline{P}_\sigma$. In particular f_σ is essentially active in the sense of Scholtes [24, Chap. 4.1] at all points in $\overline{P}_\sigma$ provided σ is open. Whether or not it is essentially active somewhere outside of $\overline{P}_\sigma$ is irrelevant and needs not be tested. To conform with the general concepts of piecewise smooth functions we may restrict f_σ to some open neighborhood of $\overline{P}_\sigma$ such that it cannot be essentially active outside P_σ. The corresponding signature vectors are given by

$$\mathcal{E} = \{\sigma \in \{-1, 0, 1\}^s : P_\sigma \text{ open}\}.$$

For all $\sigma \in \mathcal{E}$ the vector g_σ defined above represents the gradient of f restricted to P_σ, which reduces to g in the smooth case. Generally we would expect the polyhedral decomposition of the piecewise linearization to contain fewer open P_σ than the decomposition of the domain of the original function into essential smooth pieces.

Example 4.2. *For the original nonlinear function of Example 4.1, there are five open polyhedra. The piecewise linearization at the origin $\mathring{x} = 0$ contains two polyhedra as sketched in Fig. 5a one of which is critical but open. For the piecewise linearization* (10) *and the corresponing abs-normal form* (15)*, at the point $\mathring{x} = (1, 1)$, we obtain the decomposition as depicted in Fig. 5b. As can be seen, it has the four open signature vectors $\sigma = (\pm 1, \pm 1)$, which are all noncritical. Furthermore, one can observe that $f_{(-1,-1)}(x) = f_{(1,-1)}(x)$ for all $x \in \overline{P}_{(-1,-1)} \cup \overline{P}_{(1,-1)}$. Notice that this union is nonconvex so that handling the polyhedra $P_{(-1,1)}$ and $P_{(1,1)}$ as defined by the abs-normal form separately makes a lot of sense.*

Generally, we will describe the polyhedral structure primarily in terms of the signature vectors σ. They have a partial order, which is nicely reflected in the corresponding polyhedra as follows.



(a) Base point $\breve{x} = 0$ (b) Base point $\mathring{x} = (1, 1)$

Figure 5: Decomposition of the domain for the PL function (10)

Proposition 4.3 (Polyhedral structure in terms of signature vectors)**.**

(i) The signature vectors are partially ordered by the precedence relation

$$\sigma \preceq \tilde{\sigma} \;:\Longleftrightarrow\; \sigma_i^2 \leq \tilde{\sigma}_i \, \sigma_i \quad \textit{for } 1 \leq i \leq s. \tag{20}$$

(ii) The closure $\bar{P}_\sigma$ of any P_σ is contained in the extended closure

$$\hat{P}_\sigma \equiv \{x \in \mathbb{R}^n : \sigma(x) \preceq \sigma\} \;\supset\; \bar{P}_\sigma \tag{21}$$

with equality holding unless $P_\sigma = \emptyset$.

(iii) The essential signatures $\mathcal{E}$ are exactly the maximal elements amongst all nonempty signatures, i.e.

$$\mathcal{E} \ni \sigma \prec \tilde{\sigma} \;\;\Longrightarrow\;\; P_{\tilde{\sigma}} = \emptyset \quad \textit{and} \quad \hat{P}_\sigma = \hat{P}_{\tilde{\sigma}}$$

we will call such $\tilde{\sigma}$ extended essential.

(iv) For any two signatures σ and $\tilde{\sigma}$ we have the equivalence

$$\hat{P}_\sigma \subset \hat{P}_{\tilde{\sigma}} \;\;\Longleftrightarrow\;\; \sigma \preceq \tilde{\sigma}.$$

(v) Each polyhedron intersects only the extended closures of its successors

$$P_\sigma \cap \hat{P}_{\tilde{\sigma}} \neq \emptyset \;\;\Longrightarrow\;\; \sigma \preceq \tilde{\sigma}.$$

(vi) The closures of the essential polyhedra form a polyhedral decomposition in that

$$\bigcup_{\sigma\in\mathcal{E}} \hat{P}_\sigma \;=\; \mathbb{R}^n.$$

Proof.
(i): The relationship requires for each i that $\sigma_i = 0$ if $\tilde{\sigma}_i = 0$ and otherwise $\sigma_i = \tilde{\sigma}_i$ or $\sigma_i = 0$. Hence, σ is componentwise closer to the zero vector than $\tilde{\sigma}$. Obviously that is a transitive relation.
(ii): Here, we require for each i that $z_i(x) = 0$ if $\sigma_i = 0$ and otherwise that $\sigma_i\, z_i(x) \geq 0$. It can be easily checked that these continuous equalities are satisfied on a closed convex polyhedron $\hat{P}_\sigma \subset \mathbb{R}^n$, which does of course contain P_σ. Suppose now that $P_\sigma \neq \emptyset$ and $\hat{P}_\sigma \setminus P_\sigma$ contains a point x that is not in the closure $\bar{P}_\sigma$. Then we must have for some index i that $z_i(x) = 0 \neq \sigma_i$ and the same must be true on a relatively open neighborhood also contained in $\hat{P}_\sigma \setminus P_\sigma$. That would require ∇z_i to be orthogonal to the tangent space of P_σ which implies that $z_i = 0$ throughout P_σ and $\hat{P}_\sigma$, which contradicts the assumption $\sigma_i \neq 0$. Hence, we must have $\hat{P}_\sigma = \bar{P}_\sigma$. Emptiness of P_σ and thus its closure can arise as in Example 4.1 for the piecewise linearization at the origin. Here all three P_σ with $\sigma = (\sigma_1, \sigma_2) \neq (0, 0)$ are empty but their extended closures are identically equal to $\hat{P}_\sigma = \mathbb{R}^2$.
(iii) Since $\sigma \prec \tilde{\sigma}$ there must be a minimal index i such that $\sigma_i = 0 \neq \tilde{\sigma}_i$. Until then we must have $\sigma_j = \tilde{\sigma}_j$ for all $j < i$. Hence we must have at all x in the open set P_σ that $z_i = 0$ and thus $\nabla z_i = 0$. Consequently, there can be no point $\tilde{x}$ whose signature $\tilde{\sigma}$ agrees with σ up to the $(i-1)$-st component but then has $\tilde{\sigma}_i \neq 0$.
(iv): Obviously P_σ is always contained in its extended closure, which certainly is contained in $\hat{P}_{\tilde{\sigma}}$ if and only if $\sigma \preceq \tilde{\sigma}$ since the extended closures defined in (21) are certainly monotonic with respect to the signature vector ordering.
(v): Assume that there exists $x \in \mathbb{R}^n$ with $x \in P_\sigma \cap \hat{P}_{\tilde{\sigma}}$. It follows from $x \in P_\sigma$ that $\sigma(x) = \sigma$. Furthermore, one obtains from $x \in \hat{P}_{\tilde{\sigma}}$ that $\sigma(x) \preceq \tilde{\sigma}$. Therefore, one has $\sigma = \sigma(x) \preceq \tilde{\sigma}$.
(vi): If this was not true there would have to be an open domain not contained in any of the open polyhedra, which is a contradiction to the definition of the polyhedra in (16). □

Obviously, a gradient g_σ is very easy to calculate for any given open σ. To find for a given x some open σ with the closure $\overline{P}_\sigma$ containing x one may use the following trick, which we will call *polynomial escape*. Due to piecewise linearity, the complement $\mathcal{C}$ of all open P_σ is contained in the union of finitely many

hyperplanes. Hence, no polynomial path of the form

$$x(t) \equiv \sum_{i=1}^{n} e_i t^i \quad \text{with} \quad \det[e_1, e_2, \ldots, e_n] \neq 0 \quad \text{for} \quad e_i \in \mathbb{R}^n$$

can be contained in $\mathcal{C}$. In other words, we find for some σ and $\bar{t} > 0$ that $x(t) \in P_\sigma$ for all $t \in (0, \bar{t})$. The corresponding open σ can be computed by some sort of lexicographic differentiation as introduced by Nesterov [23] and described in a little more detail in [11]. There it is also shown that any such g_σ is in fact a generalized gradient of the underlying nonlinear function, if f was obtained by piecewise linearization.

By suitable selecting $e_1 = d \neq 0$ one can make sure that the generalized gradient obtained is active in a cone containing the given direction d at least in its closure. Then we may set $g(x, d) = g_\sigma$ with the properties of a directionally active gradient discussed in Section 2. A maximal bundle strategy would be to keep all the g_σ and γ_σ with their respective essential signature $\sigma \in \mathcal{E}$ in memory. In fact for the theory we will assume at first that they are all known. As a consequence of the last proposition we find:

Proposition 4.4 (Limiting gradient sets and tangent spaces)**.**

(i) At all x contained in a given P_σ we have the same limiting gradient set

$$\partial^L f(x) = \partial^L f(P_\sigma) \equiv \{g_{\tilde{\sigma}} : \sigma \preceq \tilde{\sigma} \in \mathcal{E}\}. \tag{22}$$

(ii) The extended closure of P_σ is the coincidence set of all essential $f_{\tilde{\sigma}}$ with $\tilde{\sigma} \succeq \sigma$, i.e.,

$$\hat{P}_\sigma = \{x \in \mathbb{R}^n : f(x) = f_{\tilde{\sigma}}(x) \ \text{ if } \ \sigma \preceq \tilde{\sigma} \in \mathcal{E}\}. \tag{23}$$

(iii) The tangent spaces $T(P_\sigma)$ of P_σ and $T(\partial^L f(P_\sigma))$ of $\partial^L f(P_\sigma)$ are orthogonal complements, i.e.,

$$x + v \in P_\sigma \ \text{ for } \ 0 \approx v \in \mathbb{R}^n \quad \Longleftrightarrow \quad (g - \tilde{g})^\top v = 0 \ \text{ if } \ g, \tilde{g} \in \partial^L f(P_\sigma),$$

where $x \in P_\sigma$ may be any fixed point.

Proof. (i): The assertion follows from the definition of the limiting gradients in combination with Prop. 4.3, (iv) and (v).
(ii): First assume that $\sigma \in \mathcal{E}$. Because of the definition of the precedence relation $\preceq$ in (20) one has for every $\tilde{\sigma} \in \mathcal{E}$ with $\sigma \preceq \tilde{\sigma}$ that $\sigma = \tilde{\sigma}$. Therefore one obtains

$$\{x \in \mathbb{R}^n : f(x) = f_{\tilde{\sigma}}(x) \ \text{ if } \ \sigma \preceq \tilde{\sigma} \in \mathcal{E}\} = \{x \in \mathbb{R}^n : f(x) = f_\sigma(x)\}.$$

Because of the continuity of f it follows for $x \in \partial P_\sigma$ that $f(x) = f_\sigma(x)$ yielding (23). Now assume that $\sigma \notin \mathcal{E}$. From Prop. 4.3, (vi), we have that there exists a collection $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_k \in \mathcal{E}$ with

$$P_\sigma \subset \bigcup_{1 \le i \le k} \hat{P}_{\tilde{\sigma}_i} \quad \text{such that } P_\sigma \cap \hat{P}_{\tilde{\sigma}_i} \neq \emptyset \text{ for } 1 \le i \le k.$$

This yields with Prop. 4.3, (v), that $\sigma \preceq \tilde{\sigma}_i$, for $1 \le i \le k$. Furthermore, since $\sigma \notin \mathcal{E}$ one knows that

$$x \in P_\sigma \cap \hat{P}_{\tilde{\sigma}_i} \quad \Rightarrow \quad x \in \partial P_{\tilde{\sigma}_i}.$$

Then, the continuity of f ensures that $f(x) = f_\sigma(x) = f_{\tilde{\sigma}_i}(x)$ yielding (23).
(iii): For $x \in P_\sigma$, it follows from (i) that $T(\partial^L f(P_\sigma)) = T(\partial^L f(x))$. Furthermore, $T(\partial^L f(x))$ is the linear space spanned by the shifted generalized gradient $\partial f(x) - g$ with $g \equiv \text{short}(0, \partial^L f(x))$. Its orthogonal complement V exists of all vectors $v \in \mathbb{R}^n$ for which

$$g^\top v = \tilde{g}^\top v \quad \text{if} \quad \tilde{g} \in \partial f(x) \Leftrightarrow \gamma_\sigma + g^\top v = \gamma_\sigma + \tilde{g}^\top v \quad \text{if} \quad \tilde{g} \in \partial f(x).$$

This condition is equivalent to $f_\sigma(x+v) - f_\sigma(x) = f(x+v) - f(x)$. In other words if f_σ is active at x this also applies to all $x + v$, which means that $V = T(P_\sigma)$ is indeed the tangent space of P_σ proving (iii). □

5 Minimizing a PL function with proximal term

Piecewise linear models of general objective functions may be unbounded. This can be easily seen for $f(x) = x^2$ where one obtains at $\mathring{x} = 1$ the local PL model $\Delta f(\mathring{x}; x) = 1 + 2x$. For this reason, we incorporate a so-called proximal term to ensure the boundedness of the considered objective function. That is, we consider the problem of minimizing a function of the form

$$\hat{f}(x) \equiv f(x) + \frac{q}{2}\|x - x_0\|^2 \quad \text{with} \quad q \ge 0, \tag{24}$$

where $f(x)$ is assumed to be PL and represented in abs-normal form. The vector x_0 may represent a point at which the local PL model is generated. Throughout this section, x_0 will be constant so that we may set it without loss of generality to $x_0 = 0$, which may require an adjustment in the constant vector c of the abs-normal representation (14).

We are mostly interested in the case $q > 0$ but will still cover the exactly piecewise linear case $q = 0$. Let us firstly notice some fairly obvious properties using again the notation short() as defined in (4).

Lemma 5.1 (Basic Properties)**.**

(i) As $\partial^L \hat{f}(x) = \partial^L f(x) + q\,x$ *we have*

$$short(0, -\partial^L \hat{f}(x)) = short(q\,x, -\partial^L f(x)).$$

(ii) The function $\hat{f}$ *attains a global minimum whenever it is bounded below, which must hold if* $q > 0$.

(iii) The function $\hat{f}$ *is globally convex if and only if this holds for the PL part* f.

(iv) If $q > 0$ *all first order minimal points* x_* *of* $\hat{f}$ *are isolated local minima. This implies in the convex case the uniqueness of the global minimizer* x_*.

Proof. (i): Follows from the differentiation of $\hat{f}$ in (24) and the definition of short() in (4).
(ii): Consider a sequence $\{x_k\} \subset \mathbb{R}^n$ such that

$$-\infty < \inf_{x \in \mathbb{R}^n} \hat{f}(x) = \lim_{k \to \infty} \hat{f}(x_k).$$

Since there are only finitely many polyhedra we may assume w.o.l.g. that all elements of the infimizing sequence belong to some P_σ so that

$$\hat{f}(x_k) = f_\sigma(x_k) + g_\sigma^\top x_k + q||x_k||^2/2.$$

If $q = 0$ we can consider the minimization of f over the closed polyhedron $\overline{P}_\sigma$ as an LP. For LPs it is well known that feasibility and boundedness implies the existence of an optimal solution which is of course global. If $q > 0$ then the x_k must be bounded and have a cluster point where $\hat{f}$ attains the minimal value.
(iii): The penalty term is convex so only the PL part f can destroy the convexity of $\hat{f}$. Assume that $\hat{f}$ is globally convex and consider an arbitrary point $\bar{x}$ where $\hat{f}(\bar{x}) = f(\bar{x}) + q||\bar{x}||^2/2$ has the subgradient $\bar{g}$. That implies for all x

$$\begin{array}{rcl} \bar{g}^\top(x - \bar{x}) & \leq & f(x) - f(\bar{x}) + q[||x||^2 - ||\bar{x}||^2]/2 \\ & = & f(x) - f(\bar{x}) + q(x - \bar{x})^\top(x + \bar{x})/2 \implies \\ -q||x - \bar{x}||^2/2 & \leq & f(x) - f(\bar{x}) - (\bar{g} - q\bar{x})^\top(x - \bar{x}). \end{array}$$

The function on the right hand side is piecewise linear and zero at $x = \bar{x}$. It must be in some neighborhood of $\bar{x}$ nonnegative, because if it was negative that would have to be of first order. Hence, $f(x)$ has at $\bar{x}$ the local subgradient $\bar{g} - q\,\bar{x}$. That implies the convexity of f by the following argument. Suppose f was not convex along a line from some $\bar{x}$ to some $\mathring{x}$. Then its restriction to the line would have to be nonconvex in the neighborhood of some kink, i.e., the slope would not be

monotonic. That is excluded by the existence of a local supporting hyper plane. (iv): From the first order necessary optimality condition for x_* one obtains $f'(x_*; v) + qx_*^\top v \geq 0$ for all v. Since $\hat{f}$ is directionally quadratic this implies for fixed $v \neq 0$ and variable $t > 0$ by an Taylor expansion that

$$\begin{aligned}\hat{f}(x_* + tv) &= f(x_*) + q||x_*||^2/2 + f'(x_*; v) + qx_*^\top v + qt||v||^2/2 \\ &\geq f(x_*) + qt||v||^2/2 > f(x_*).\end{aligned}$$

If f is convex then x_* is a unique global minimizer. □

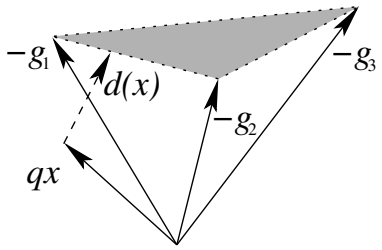


Figure 6: Direction of safe descent $d = d(x) = \text{short}(q\,x, -\partial^L f(x))$

The geometry of the first assertion (i) is depicted in Fig. 6 for the simple case $\partial^L f(x) = \{g_1, g_2, g_3\}$. The convex hull $\text{conv}(\partial^L f(x))$ is illustrated by the area shaded in gray. The safe descent $d = d(x)$ for $\hat{f}$ at x is given as the projection of qx onto this convex set. The step computation of Algo. 3.1 can be extended for this situation of a PL function with proximal term in the following way:

Algorithm 5.2 (Step Computation II)**.**

ComputeStep(x**,** q**,** G**)** // *Precondition:* $x \in \mathbb{R}^n$, $q \geq 0$, $\emptyset \neq G \subset \partial^L f(x)$
repeat
  { $d = -short(qx, G)$
    $g = g(x; d)$
    $G = G \cup \{g\}$
  }
until $(g + q\,x)^\top d \leq -\|d\|^2$
eliminate all $\tilde{g} \in G$ *with* $\tilde{g}^\top d \neq g^\top d$
return d, G

Since the proximal term results only in a linear shift of the gradient, the finite termination of Algo. 5.2 can be shown with exactly the same arguments used in

the proof of Prop. 3.2 to establish the finite termination of Algo. 3.1. We obtain the *generalized steepest descent property* in the form

$$\hat{f}'(x;d) = -\hat{g}(x;d)^\top d \;\leq\; -||d||^2 \;<\; 0. \tag{25}$$

Now let us again begin by looking at the convex case. As we noted in the introduction it was stated in [17] that for the initial condition $x(0) = x_0 = 0$ there exists a unique solution $x(t)$ with $t \in [0, \infty)$ to the differential equation

$$D_+x(t) = d(x(t)) \equiv \text{short}(q\,x(t), -\partial^L f(x(t))). \tag{26}$$

Here we have used the first assertion of the previous Lemma 5.1 to express the right hand side directly in terms of f or rather its limiting gradient set.
From this fundamental result one can derive as in [17, Chap. VIII, Theorem 3.4.1 and Corollary 3.4.2] the following implications:

Corollary 5.3 (Convergence properties in convex case)**.**
Assume that the PL function f is convex. Then:
(i) The function value $\hat{f}(x(t))$ satisfies

$$D_+\hat{f}(x(t)) = -\|d(x(t))\|^2 \leq 0. \tag{27}$$

Moreover $\hat{f}(x(t))$ is convex as $\|d(x(t))\|^2$ decreases monotonically.
(ii) If $\hat{f}$ is bounded below we have for any stationary point $z \in \mathbb{R}^n$ of $\hat{f}$

$$D_+\left(\tfrac{1}{2}\|x(t) - z\|^2\right) \leq 0,$$

where strict inequality holds if $q > 0$ and $x(t) \neq z$.
(iii) There exists a stationary limit

$$x_* = \lim_{t\to\infty} x(t) \quad \textit{with} \quad 0 \in \partial\hat{f}(x_*).$$

These very interesting properties hold for arbitrary convex $\hat{f}$. From our point of view convergence to a stationary point is not entirely satisfactory since we would really like that $x_* = x(t_*)$ for some finite $t_* < \infty$, and furthermore we wish to make sure that there is no Zeno effect. Finiteness must occur when $d(x(t))$ is bounded away from zero, but that does not even hold for the trivial problem

$$D_+x(t) = -\partial(q\,x(t)^2/2) = -q\,x(t) \quad \text{with} \quad x(0) = x_0 \equiv 1$$

It has the solution $x(t) = \exp(-q\,t)$ and thus an infinitely long trajectory converging to $x_* = 0$. To remedy the situation we will have to slightly rescale the

trajectory. First let us consider the geometry of the trajectory in our specific situation.

Let us consider some particular point $x_\sigma = x(t_\sigma) \in P_\sigma$ with $\sigma = \sigma(x)$ along the trajectory defined by (26). Then the question whether the steepest descent trajectory stays at least for some nearby values of t within P_σ and what it looks like can be answered as follows.

Theorem 5.4 (Invariance)**.**
If $f : \mathbb{R}^n \to \mathbb{R}$ is PL but not necessarily convex then one has:

(i) The polyhedron P_σ is invariant with respect to $\hat{f}$ in that the direction $d(x)$ belongs at all $x \in P_\sigma$ to the tangent space $T(P_\sigma)$ if and only if for one and thus all $x \in P_\sigma$

$$q\,x \in \partial f(P_\sigma) + T(P_\sigma).$$

(ii) For an invariant P_σ, $\hat{x} \in P_\sigma$, and $\hat{d} \equiv short(q\,\hat{x}, -\partial^L f(P_\sigma))$ the trajectory is given by

$$x(\hat{t}+t) = \begin{cases} \tilde{x}((1-\exp(-q\,t))/q) & \text{if } q > 0 \\ \tilde{x}(t) & \text{if } q = 0 \end{cases},$$

where

$$\tilde{x}(\tau) = \hat{x} + \tau\hat{d} \quad \text{and} \quad d(\tilde{x}(\tau)) = (1 - q\,\tau)\,\hat{d} \quad \text{for} \quad 0 \lesssim \tau \in \mathbb{R}. \tag{28}$$

(iii) If f is convex then at any $x \in \mathbb{R}^n$ there exists a positive bound $\hat{\tau} = \hat{\tau}(x) \leq 1/q$ such that the points $\{x + \tau\,d(x), 0 < \tau < \hat{\tau}(x)\}$ belong to an invariant polyhedron P_σ with

$$d(x + \tau\,d(x)) = (1 - q\tau)d(x),$$

i.e. $d(x + \tau\,d(x)) \parallel d(x) \in T(P_\sigma)$. Moreover, we have $\hat{\tau} = 1/q$ or $\hat{x} = x + \hat{\tau}d(x) \in P_{\tilde{\sigma}}$ for some $\tilde{\sigma} \prec \sigma$ and

$$d(\hat{x}) = (1 - q\,\hat{\tau})d(x) \quad \text{or} \quad \|d(\hat{x})\| < (1 - q\,\hat{\tau})\|d(x)\|. \tag{29}$$

Proof. (i): Let P_σ be an invariant polyhedron, i.e., $d(x)$ belongs at some $x \in P_\sigma$ to the tangent space $T(P_\sigma)$. Due to Prop. 4.4 (iii), one has $d(x) \in T(\partial^L f(P_\sigma))^\perp$. This is equivalent to the existence of $g \in \partial^L f(P_\sigma)$ such that $d(x) = g - qx$ and

$$g - qx \in T(\partial^L f(P_\sigma))^\perp \;\Leftrightarrow\; qx \in \partial^L f(P_\sigma) + T(\partial^L f(P_\sigma))^\perp \subset \partial f(P_\sigma) + T(P_\sigma)$$

proving (i).
(ii): By Prop. 4.4 (iii), we have for any $x \in P_\sigma$ and $x+v \in P_\sigma$ that $q(x-v)-qx = qv \in T(P_\sigma)$ is orthogonal to the tangent space of $\partial^L f(x) = \partial^L f(x+v)$. Therefore, one obtains

$$d(x+v) = \text{short}(q(x+v), -\partial^L f(x+v)) = \text{short}(qx, -\partial^L f(x)) + qv$$

and hence we have $d(x+v) \in T(P_\sigma) \iff d(x) \in T(P_\sigma)$. For $v = \tau d(x)$ with τ small enough it follows that $x + \tau d(x) \in P_\sigma$ and

$$\begin{aligned} d(x+\tau d(x)) &= -\text{short}(q(x+\tau d(x)), \partial^L f(x+\tau d(x))) \\ &= -\text{short}(qx, \partial^L f(x)) - q\tau d(x) = (1-q\tau)d(x). \end{aligned}$$

To prove the assertion we then can use (i) to obtain with the last equation

$$d(\tilde{x}(\tau)) = -\text{short}(q(\hat{x}+\tau\hat{d}), \partial^L f(\hat{x})) = -\text{short}(q\hat{x}, \partial^L f(\hat{x})) - q\tau d = (1-q\,\tau)d.$$

Hence, the constant tangent $\hat{d}$ of the straight line $\tilde{x}(\tau)$ equals for $\tau < 1/q$ indeed $1/(1-q\,\tau)$ times the steepest descent direction $d(\tilde{x}(\tau))$ at those points and is therefore just a reparametrization of $x(t)$. For $q = 0$ we have $t = \tau$ and $d(\tilde{x}(\tau)) = \hat{d} = d(x(t))$ as desired. For $q > 0$ we have $\tau = (1 - \exp(-q\,t))/q$. Differentiation yields

$$\frac{d}{dt}x(\hat{t}+t) = \frac{d}{d\tau}\tilde{x}(\tau)\frac{d}{dt}\tau = \hat{d}\,\exp(-qt) = (1-q\,\tau)\hat{d} = d(x(\hat{t}+t)).$$

(iii) Due to the outer semicontinuity of the subdifferential, one obtains for sufficiently small τ that $\partial^L f(x+\tau d(x)) \subset \partial^L f(x)$. There exists one directionally active gradient $g(x) \in \partial f(x)$ such that

$$(g(x)+qx)^\top d(x) = -\|d(x)\|^2 \geq (g+qx)^\top d(x) \quad \forall g \in \partial f(x).$$

Since f is assumed to be convex, all elements in $\partial^L f(x)$ that contribute nontrivally to $d(x)$ must be contained also in $\partial^L f(x+\tau d(x))$. Furthermore, one has that $\{x + \tau\, d(x) \,:\, 0 < \tau < \hat{\tau}(x)\} \subset P_\sigma$ due to the piecewise linearity of f. These observations yield

$$\partial^L f(P_\sigma) = \{g \in \partial^L f(x) \,:\, g^\top d(x) = f'(x;d)\} = \text{argmin}\{g^\top d(x) \,:\, g \in \partial^L f(x)\}.$$

Hence, in going from $\partial^L \hat{f}(x)$ to its subset $\partial^L \hat{f}(x+\tau d(x))$, we only loose elements g that are further away from x than $x+\tau d(x)$ and will therefore also play no role in determining $d(x+\tau d(x))$.

For $\hat{x} = x + \hat{\tau}d(x)$ assume first that $\partial^L f(x) = \partial^L f(\hat{x})$. Then as in (ii) one obtains directly $d(\hat{x}) = (1 - q\hat{\tau})d(x)$, i.e., the left-hand side of (29). Second assume that $\partial^L f(x) \neq \partial^L f(\hat{x})$. Then, it may happen that the new elements lie in the convex hull spanned by the elements of $\partial^L f(x)$. Then, once more we obtain as above $d(\hat{x}) = (1 - q\hat{\tau})d(x)$. Otherwise, we can conclude from the assumed convexity of f and (28) as shown in (ii) that $(1 - q\hat{\tau})\|d(x)\|$ is an upper bound for $\|d(\hat{x})\|$. Using $\hat{\tau} \leq 1/q$ this proves the right-hand side of (29). □

Obviously all open polyhedra P_σ must be invariant since their tangent space is the whole of $\mathbb{R}^n$. There $d(x)$ is simply $qx - \partial f(x)$ with $\partial f(x) = \{g_\sigma\}$ being a singleton formed by the proper gradient. If $\tilde{x}(\tau)$ as defined in Theo. 5.4 (ii) for a given $\hat{x}$ stays within any P_σ for all $0 \leq \tau < \hat{\tau} = 1/q < \infty$ we reach a stationary point $x_* = \tilde{x}(\hat{\tau})$ belonging to the closure of P_σ. If $q = 0$ we must have $d = 0$ and thus $0 \in \partial f(\hat{x}) = \partial \hat{f}(\hat{x})$ since otherwise $\hat{f} = f$ would be unbounded below, contrary to our general assumption. Then we have $t_* = \hat{\tau}$ which we may always use when $d = 0$ even if $q > 0$. Using the abs-normal form (14) one can write a subroutine using for example a bisection strategy in combination with algorithmic differentiation that computes the critical multiplier defined in (*iii*) of the Theorem 5.4. Therefore, we state here only a very general version:

Algorithm 5.5 (Computation of Critical Multiplier)**.**

> **CritMult**(x, d, $\hat{\tau}$) *// Precondition:* $x \in \mathbb{R}^n$, $0 \neq d \in \mathbb{R}^n$
>
> $\sigma = \sigma(x)$
>
> $\hat{\tau} = \max\{\tau | \sigma \preceq \sigma(x + \tilde{\tau}d)$ *for all* $0 < \tilde{\tau} < \tau$ *and* $\sigma(x + \tilde{\tau}d) \not\preceq \sigma(x + \tau d)\}$
>
> *return* $\hat{\tau}$

Mode details on the realisation of this algorithm can be found in [7]. This provides the line-search in the following Algorithm 5.6. It effectively generalizes Algorithm 3.4.6 in [17] to the situation with a proximal term. It is also well defined in the non-convex case, but then it is still not clear whether the quality of the steps is good enough to ensure global convergence.

It is important to recall that for a convex PL function f, and an arbitrary chosen d as input, Algo. 5.2, i.e., **ComputeStep**(x, q, G), returns exactly $d(x)$ and therefore the steepest descent direction, for which Theo. 5.4, (iii), holds. Moreover if the step stays within the closure of the current polyhedron the next iterate will be a solution and the stopping criterion will be satisfied on the next iteration due to d being zero. However, if f is not convex then the routine **ComputeStep**(x, q, G) just returns a safe descent direction d.

Now, we consider the global convergence of the algorithm in the convex case.

Algorithm 5.6 (True Descent Algorithm)**.**

PLmin(x**,** q**)** *// Precondition:* $x \in \mathbb{R}^n$, $q \geq 0$
$d =$ *rand()*
$G = \emptyset$
do
{ $g = g(x; d)$, $G = G \cup \{g\}$
$d =$ **ComputeStep(**x**,** q**,** G**)**
if $d = 0$*: stop*
CritMult(x**,** d**,** $\hat{\tau}$**)**
$x = x + \hat{\tau} d$
Eliminate all $g \in G$ *with* $\sigma(g) \not\succ \sigma(x)$
}

Theorem 5.7 (Convergence in the convex case)**.** *Suppose* f *is PL and with* $q \geq 0$ *and* $x_0 \in \mathbb{R}^n$ *fixed that* $\hat{f}(x) = f(x) + \frac{q}{2}\|x - x_0\|^2$ *is convex. Then Algorithm 5.6 generates a sequence of iterates* x_k *such that*

$$\lim_{k\to\infty} \hat{f}(x_k) = \hat{f}_* \equiv \inf_{x\in\mathbb{R}^n} \hat{f}(x) \geq -\infty$$

with $x_* = x_k$ *a minimizer of* $\hat{f}$ *for all large* k *if* $\hat{f}$ *is bounded below.*

Proof. Again we may assume without loss of generality that $x_0 = 0$ and $f(x_0) = 0$. If the monotonically falling values $\hat{f}(x_k)$ are not bounded below we must have $\hat{f}_* = -\infty$ and nothing remains to be shown. Otherwise, it follows that

$$\hat{f}_* \leq \hat{f}(x_{k+1}) - \hat{f}(x_k) = -\hat{\tau}_k g_k^\top d_k/2 \leq -\hat{\tau}_k \|d_k\|^2/2. \tag{30}$$

Now let us suppose first that the $\|d(x_k)\|$ are not bounded away from zero. Then either $q = 0$ in which case d_k must reach 0 exactly after finitely many steps or $q > 0$ so that the x_k are bounded and must have a stationary cluster point x_*. In either case the stationary point must be by assumption of convexity globally minimal. Moreover, even in the second case the sequence must reach a first point x_{k-1} in one of the finitely many polyhedra P_σ whose closure $\overline{P}_\sigma$ contains x_*. Since $\hat{f}$ is convex, the next iterates can not leave $\overline{P}_\sigma$ anymore. Then, due to the employed line-search, x_* must be reached in a finite number of steps proving the assertion. That leaves us with the possibility that

$$\inf_{k\in\mathbb{N}} \|d_k\| = \lim_{k\to\infty} \|d_k\| > 0.$$

Here the first equality follows from the fact that the $\|d_k\|$ decline monotonically as a consequence of the assumed convexity of $\hat{f}$. Then it follows by summation of the above telescoping series $\hat{f}(x_{k+1}) - \hat{f}(x_k)$ and the boundedness of $\hat{f}_*$ from (30) that the $\hat{\tau}_k$ and thus the steps lengths $\hat{\tau}_k\|d_k\|$ are summable. Then, the x_k must have a unique limit point x_* and the $\hat{\tau}_k$ must converge to zero.

Let $(x_{k_j})_{j\in\mathbb{N}}$ denote the subsequence of $(x_k)_{k\in\mathbb{N}}$ that belongs to one of the finitely many polyhedra P_{σ_i}, $1 \le i \le l$, whose closure contains x_*. Then we must have due to the continuity of the projection operator

$$\lim_{j\to\infty} d(x_{k_j}) = \lim_{j\to\infty} \mathrm{short}(q\, x_{k_j}, \partial^L(P_{\sigma_i})) = d_{\sigma_i} \equiv \mathrm{short}(q\, x_*, \partial^L(P_{\sigma_i})).$$

Hence, the monotonically declining norms $\|d_{k_j}\|$ must converge to exactly one particular value $\|d_\sigma\|$ and after a possible renumbering all late x_k must belong to a subset of polyhedra $\bigcup P_{\sigma_i}$, $1 \le i \le \hat{l} \le l$, for which $\|d_{\sigma_i}\| = \|d_\sigma\|$. To derive a contradiction, first assume that there exists a $\bar{d}$ such that $\bar{d} = d_{\sigma_i}$ for all $1 \le i \le \hat{l}$. The definition of the step multiplier $\hat{\tau}_k$ in Theo. 5.4 (ii) ensures that all iterates x_k lie on a "kink", i.e., for each k there exists $i_k, \hat{i}_k \in \{1, \ldots, \hat{l}\}, i_k \neq \hat{i}_k$, with $x \in \overline{P}_{\sigma_{i_k}} \cap \overline{P}_{\sigma_{\hat{i}_k}}$. Since there are infinitely many iterates there must be infinitely many kinks, and therefore also infinitely many polyhedra, along the direction $\bar{d}$. This is a contradiction to the property that f is a PL function. Hence, there must exists at least one $\hat{k}$ such that $\|d_{\hat{k}}\| = \|d_{\hat{k}+1}\|$ but $d_{\hat{k}} \neq d_{\hat{k}+1}$. Then, $\partial^L f(x_{\hat{k}+1})$ contains all gradients that contribute to $d_{\hat{k}}$ and $d_{\hat{k}+1}$ such that $\tilde{d} = \frac{1}{2}(d_{\hat{k}} + d_{\hat{k}+1})$ represents a convex combination of gradients contained in $\partial^L f(x_{\hat{k}+1})$ with

$$\|\tilde{d}\|^2 = \|\frac{1}{2}(d_{\hat{k}} + d_{\hat{k}+1})\|^2 < \|d_{\hat{k}}\| = \|d_{\hat{k}+1}\|.$$

This yields a contradiction to the choice of $d_{\hat{k}+1}$ as steepest descent direction. Therefore, it is shown that the $\|d(x_k)\|$ can not be bounded away from zero yielding convergence of the iterates as shown above. □

The result above is theoretically quite satisfactory. However, its implementation in the presence of rounding errors is rather challenging. First and foremost one must keep track of the currently active constraints, which manifest themselves in zeros of the signature vector σ describing the polyhedron P_σ that the current iterate belongs to. Then the steps d must be computed accordingly. Similarly delicate is the management of the bundle G, whose elements should be purged if they no longer belong to the limiting Jacobian as characterized in Proposition 4.4. So far we have tested that in the routine **ComputeStep(**x**,** q**,** G**)** indirectly, which is correct for the convex case.

	f^*	$\#f$	$\#\nabla f$	# QP	iter
plmin	-100	21	24	12	5
hanso	-100	7	7	–	3 BFGS
MPBNGC	-100	13	13	–	8

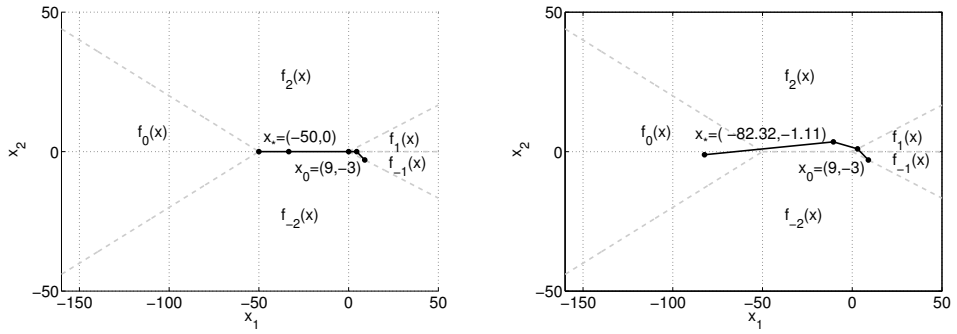
Table 1: Function values reached and evaluation counts for Hiriart-Urruty/Lemaréchal example

6 Experimental Verification

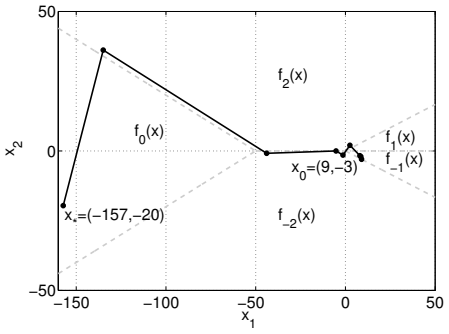
To illustrate the behavior of the new optimization approach, we coded a first version plmin of Algorithm 5.6, applied it to several standard test problems and compared the convergence behaviour with the proximal bundle method MPBNGC [21] as well as the quasi-Newton method hanso that is adapted for the nonsmooth case [20]. Both algorithms use functions values and gradient information for the optimization but do not have the possibility to exploit the abs-normal form. We used the proposed values of the parameters for both packages. Furthermore, for MPBNGC we set the bundle size equal to the number of variables for a fair comparison.

Example by Hiriart-Urruty and Lemaréchal

Applying the true steepest descent algorithm plmin to the objective function (1) proposed by Hiriart-Urruty and Lemaréchal, we reach an optimum after four iterations as shown in Fig. 7a. The three iterates needed by hanso are illustrated in Fig. 7b. Finally, Fig. 7c shows the convergence history of MPBNGC. All three methods reach an optimal point. The number of function evaluations, gradient evaluations and solves of a QP to reach an optimal point are reported in Tab. 1. As can be seen, for this example our optimization routine needs more evaluations of the function and its gradients. However, as illustrated in Fig. 7, the iterates are chosen in a systematical way. As can be seen from the next examples, this leads to a dramatic reduction of the number of function and gradient evaluations for more complex examples.



(a) Iterates of plmin, i.e., Algorithm 5.6 (b) Iterates of hanso



(c) Iterates of MPBNGC

Figure 7: Optimization history for the Hiriart-Urruty/Lemaréchal example

L1 hilb

We also considered the scalable L1hilb function [29]

$$f : \mathbb{R}^n \mapsto \mathbb{R}, \quad f(x) = \sum_{i=1}^{n} \left| \sum_{i=1}^{n} \frac{x_j}{i+j-1} \right|. \tag{31}$$

A remarkable property of the function (31) is the appearance of gradients g_σ and $g_{\tilde{\sigma}}$ with $g_\sigma = -g_{\tilde{\sigma}}$ and $\sigma \neq \tilde{\sigma}$. The corresponding polyhedra P_σ and $P_{\tilde{\sigma}}$ only have one single point in common. Whenever both gradients g_σ, $g_{\tilde{\sigma}}$ are elements of the bundle it is possible to combine them linearly to 0. That is why it is very important in this case to eliminate elements of the bundle that do not belong to neighbouring polyhedra of the current iterate.

Again we compared our first implementation of Algo. 5.6 with hanso and MPBNGC. The results are shown in Tab. 2, where GS stands for Gradient Sampling. As can be seen, the iterations count and hence also the number of function

n		f^*	$\#f$	$\#\nabla f$	# QP	iter
2	plmin	0	21	53	11	4
	hanso	0.0275	9766	9766	–	6 BFGS + 300 GS
	MPBNGC	0.2319	99804	99804	–	1000
3	plmin	0	49	53	25	10
	hanso	0.0017	10725	10725	–	14 BFGS + 300 GS
	MPBNGC	0.8380	99803	99803	–	1000
4	plmin	0	97	100	49	18
	hanso	0.0031	10473	10473	–	7 BFGS + 300 GS
	MPBNGC	0.3888	99803	99803	–	1000
5	plmin	$1.6e-13$	189	197	95	47
	hanso	0.0046	10744	10744	–	5 BFGS + 300 GS
	MPBNGC	0.3837	5991	5991	–	1000
6	plmin	$3.1e-11$	209	216	105	79
	hanso	0.0181	11557	11557	–	5 BFGS + 300 GS
	MPBNGC	0.8323	4993	4993	–	1000

Table 2: Function values reached and evaluation counts for L1hilb example

and gradient evaluations as well as the QP solves required by plmin increase considerably with the dimension. This is due to the rather complicated structure of the polyhedra decomposition as detailed above. However, the optimal point is reached within a reasonable number of iterations. The algorithm hanso always terminates when the gradient sampling yielded a smaller value of the target function, but did not reach the optimal point. MPBNGC always terminates as the maximal number of iterations is reached without attaining the optimal point.

Goffin

As next example we consider the Goffin function [1]

$$f : \mathbb{R}^{50} \mapsto \mathbb{R}, \quad f(x) = 50 \max_{1 \le i \le 50} x_i - \sum_{i=1}^{50} x_i \, . \tag{32}$$

All three methods reach an optimal point, but only plmin and MPBNGC terminte regularily. That is, the termination criteria of hanso did not work. Therefore, we stopped the optimization process of hanso manually when the same accuracy was met, which is marked with * in Tab. 3. This tables also shows the number of

	f^*	$\#f$	$\#\nabla f$	# QP	iter
plmin	0	227	292	146	69
hanso	0	3017	3017	–	706* BFGS
MPBNGC	0	556	556	–	555

Table 3: Function values reached and evaluation counts for Goffin function

function evaluations, gradient evaluations and solves of a QP to reach an optimal point are reported in Tab. 3. As can be seen, our algorithm plmin needed the fewest iterations.

Linear Nonsmooth Rosenbrock á la Nesterov

Whereas the previous examples have convex target functions, the linear nonsmooth version of the classical Rosenbrock function proposed by Nesterov [15], i.e.,

$$f : \mathbb{R}^{50} \mapsto \mathbb{R}, \quad f(x) = \tfrac{1}{4}|x_1 - 1| + \sum_{i=1}^{n-1} |x_{i+1} - 2|x_i| + 1| \ .$$

is not convex. Here, we report results for two different starting points. The first one, i.e., $x_0 = 0 \in \mathbb{R}^n$ lies in a convex part of the function including also the optimal point. This is not the case for the second one, i.e., $x_0 = (0.5 * (-1)^i)_{i=1,\ldots,n} \in \mathbb{R}^n$. The function values reached, the number of function evaluations, gradient evaluations and solves of a QP are reported in Tab. 4 and 5. As can be seen from Tab. 5 the optimization becomes much more challenging in the nonconvex case. Here, the algorithm hanso always terminates for $n > 2$ when the gradient sampling yielded a smaller value of the target function, whereas MPBNGC terminates at a stationary point in very few iterations. Our algorithm terminates also at a stationary point in accordance with the theory presented here needing a few more iterations.

Nonlinear Nonsmooth Rosenbrock á la Nesterov

Finally, we show how a piecewise smooth optimization problem can be solved by successive piecewise linearization. That is, the minimization of PL functions presented here is used as an inner loop of an outer optimization algorithm as sketched in [11, Sect. 5.2]. For this purpose the piecewise linearization is generated

n		f^*	$\#f$	$\#\nabla f$	# QP	iter
2	plmin	0	9	10	5	2
	hanso	0	587	587	–	128 BFGS + 5 GS
	MPBNGC	0.25	2	2	–	1
3	plmin	0	17	18	9	4
	hanso	0	1420	1420	–	68 BFGS + 27 GS
	MPBNGC	0.125	2	2	–	1
4	plmin	0	41	42	21	10
	hanso	0.0625	5770	5770	–	242 BFGS + 127 GS
	MPBNGC	0.225	2	2	–	1
5	plmin	$1.36e-13$	25	26	13	6
	hanso	0.1311	8308	8308	–	238 BFGS + 209 GS
	MPBNGC	0.325	2	2	–	1

Table 4: Function values reached and evaluation counts for linear nonsmooth Rosenbrock and $x_0 = 0 \in \mathbb{R}^n$

n		f^*	$\#f$	$\#\nabla f$	# QP	iter
2	plmin	0.25	9	10	5	2
	hanso	0	274	274	–	88 BFGS
	MPBNGC	0.375	3	3	–	2
3	plmin	0.375	13	15	7	3
	hanso	0.125	12354	12354	–	479 BFGS + 300 GS
	MPBNGC	0.875	5	5	–	4
4	plmin	0.375	29	30	15	7
	hanso	0.25	14621	14621	–	1000 BFGS + 300 GS
	MPBNGC	1.375	3	3	–	2
5	plmin	0.375	45	50	25	11
	hanso	0.313	15450	15450	–	1000 BFGS + 300 GS
	MPBNGC	1.875	3	3	–	2

Table 5: Function values reached and evaluation counts for linear nonsmooth Rosenbrock and $x_0 = (0.5 * (-1)^i)_{i=1,\ldots,n} \in \mathbb{R}^n$

at each iterate of the outer iteration and the penalty parameter q is adapted correspondingly.

Nesterov suggested also a nonlinear nonsmooth version of the classical Rosenbrock function, specifically for $n = 2$

$$f(x_1, x_2) = \tfrac{1}{4}(x_1 - 1)^2 + \left|x_2 - 2\,x_1^2 + 1\right| \ .$$

At a point $\mathring{x}_1, \mathring{x}_2$ one obtains the piecewise linearization

$$\begin{aligned} f(\mathring{x}_1, \mathring{x}_2) + \Delta f(\mathring{x}_1, \mathring{x}_2; \Delta x_1, \Delta x_2) \ = \\ \tfrac{1}{4}(\mathring{x}_1 - 1)^2 + \tfrac{1}{2}(\mathring{x}_1 - 1)\Delta x_1 + \left|\mathring{x}_2 + \Delta x_2 - 2\,\mathring{x}_1^2 - 4\mathring{x}_1\Delta x_1 + 1\right| \ . \end{aligned}$$

Subtracting the right hand side from $f(\mathring{x}_1 + \Delta x_1, \mathring{x}_2 + \Delta x_2)$ and taking the absolute value we obtain the discrepancy

$$\begin{aligned} &\Big|\tfrac{1}{4}(\Delta x_1)^2 + \left|\mathring{x}_2 + \Delta x_2 - 2\,(\mathring{x}_1 + \Delta x_1)^2 + 1\right| - \left|\mathring{x}_2 + \Delta x_2 - 2\,\mathring{x}_1^2 - 4\mathring{x}_1\Delta x_1 + 1\right|\Big| \\ &\quad \le \tfrac{1}{4}(\Delta x_1)^2 + \left|2\,(\mathring{x}_1 + \Delta x_1)^2 - 2\,\mathring{x}_1^2 - 4\mathring{x}_1\Delta x_1\,\right| \ = \ q\,(\Delta x_1)^2 \quad \text{with} \quad q \ = \ \tfrac{9}{4}\,. \end{aligned}$$

Now suppose we successively minimize the convex piecewise linearization with proximal term $q[(\Delta x_1)^2 + (\Delta x_2)^2]$ defined by that maximal value of q as suggested in [11, Sect. 5.2]. There convergence has been established, so we may assume that the current point $(\mathring{x}_1, \mathring{x}_2)$ is already close to the optimal solution $x^* = (x_1^*, x_2^*) = (1, 1)$. If the next iterate $(x_1^+, x_2^+) = (\mathring{x}_1 + \Delta x_1, \mathring{x}_2 + \Delta x_2)$ did not lie on the kink of the current PL model, differentiation with respect to Δx_2 would yield the condition $\pm 1 = 2q\,\Delta x_2$ and thus $|\Delta x_2| = 0.5/q = \frac{2}{9}$. This would clearly prevent convergence so that the PL minimizer must lie on the kink. Differentiating the remaining terms with respect to Δx_1 we obtain the condition $\frac{1}{2}(\mathring{x}_1 - 1) + 2q\Delta x_1 = 0$, which yields $\Delta x_1 = (1 - \mathring{x}_1)/(4q) = (1 - \mathring{x}_1)/9$. Thus we obtain

$$(x_1^+ - 1) = (\mathring{x}_1 + \Delta x_1 - 1) = (\mathring{x}_1 - 1)(8/9)$$

which means linear convergence with the rate $8/9$ towards $x_1^* = 1$. The other component is always adjusted so that $x_2^+ = 2\,(x_1^+)^2 - 1 - 2(\Delta x_1)^2$ and thus also converges linearly towards the optimal value $x_2^* = 1$. This convergence behaviour is also illustrated in the contour plot Fig. 8 as well and in Fig. 9 showing the linear convergence and the development of the penalty factor q. These results were obtained with a preliminary implementation of the piecewise linearization approach starting with $q = 1.0$.

The reduction factor $8/9$ may not seem very impressive, but it is much better than the asymptotic rate $1 - 1/\kappa$ with $\kappa \approx 2.500$ that steepest descent achieves

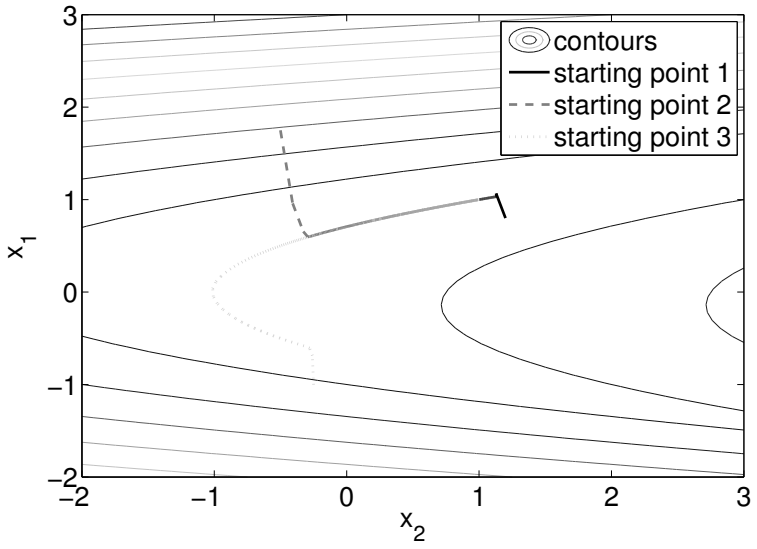


Figure 8: Contours and iterates generated by Algo. 5.6 from three starting points.

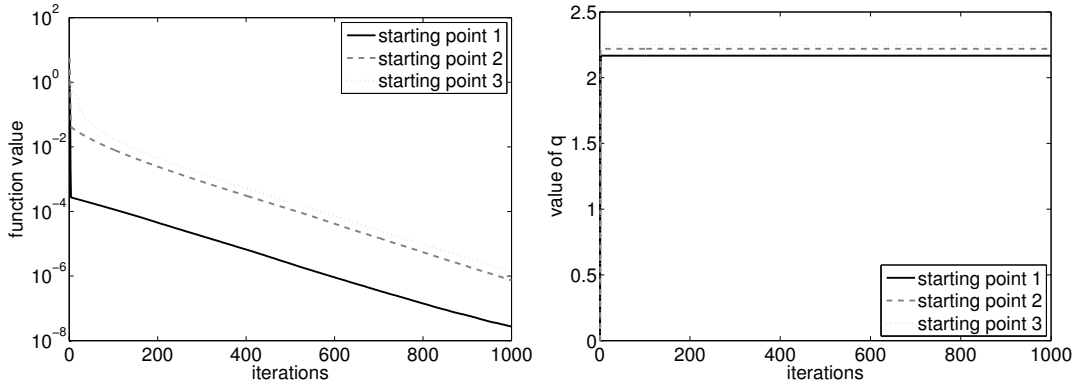


Figure 9: Function values of iterates (left) and values of q (right) for the three different starting points as above

on the smooth variant of the Rosenbrock function. Generally, in the smooth case successive piecewise linearization with a proximal term also reduces to steepest descent with a particular step size rule. Thus we cannot expect to achieve anything like a superlinear rate of convergence. That is only possible if one replaces the proximal term with a quadratic $(x - \mathring{x})^\top B(x - \mathring{x})/2$, where B approximates the Hessian of a suitable Lagrangian function. As of now that seems like rather a remote possibility and we will have to accept linear convergence at any reasonable rate.

7 Conclusion and Outlook

In this paper, we present and analyze a gray-box scenario for the optimization of composite Lipschitzian objective functions. The key ingredient is the concept of piecewise linearization obtained in the abs-normal form in an AD-like fashion. The resulting structural information provides directionally active gradients and critical step multipliers, which form the basis of the new bundle method for minimizing piecewise linear functions with a proximal term. In the convex and bounded case, the method coincides with the search trajectory analyzed in [17], and convergence in finitely many steps is guaranteed. Preliminary numerical results give a first impression of the performance of the algorithm. At least in nondegenerate cases there is the possibility to extract more information from the abs-normal form, namely to evaluate complete limiting gradients as characterized in Proposition 4.4 and to test for local convexity near stationary points. Such pieces of information are available at a reasonable cost.

As demonstrated on the nonsmooth Rosenbrock function of Nesterov, this method can serve as inner loop in a quadratic overestimation scheme for the minimization of piecewise smooth objectives. We are currently developing a convergence theory for the non-convex case. We will also provide a stable general implementation together with a comprehensive testing and comparisons with other nonsmooth optimization schemes. All this will be the subject of the forthcoming paper [14].

A natural extension of the problem considered here is the minimization of a residual $\|F(x)\|_p$ for a piecewise linear vector function $F : \mathbb{R}^n \mapsto \mathbb{R}^m$. When $p = 1$ or $p = \infty$ we have again an unconstrained PL optimization problem, but the particular structure could possibly be exploited to improve efficiency. When $p = 2$ we have a least squares problem, where the polyhedral structure is inherited from $F(x)$ but the quadratic term may jump at the interfaces. The formally well-determined case $m = n$ of piecewise linear equation solving in abs-normal form

has recently been studied in [12]. Finding a stationary point of a generalized gradient is the symmetric variant of solving an algebraic inclusion $0 \in F(x)$ where $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a convex outer semi-continuous multifunction. That more general problem and corresponding differential conclusions [2] can also be attacked via successive piecewise linearizations, though the local PL models need no longer be continuous as was assumed so far based on the framework of [11]. Here a generalization to discontinuous models would be a significant departure.

References

[1] W. Alt. *Numerische Verfahren der konvexen, nichtglatten Optimierung. Eine anwendungsorientierte Einführung.* Teubner, 2004.

[2] J.-P. Aubin and C. Arriga. *Differential inclusions. Set-valued maps and viability theory.* Springer, 1984.

[3] F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Numerical optimization. Theoretical and practical aspects. Transl. from the French. 2nd reviseded.* Springer, 2006.

[4] F. Clarke. *Optimization and Nonsmooth Analysis.* SIAM, 1990.

[5] R. Cominetti and M. Courdurier. Coupling general penalty schemes for convex programming with the steepest descent and the proximal point algorithm. *SIAM J. Optim.*, 13(3):745–765, 2002.

[6] W. de Oliveira and C. Sagastizábal. Bundle methods in the XXIst century: A birds'-eye view. *Pesquisa Operacional*, 34(3):647–670, 2014.

[7] S. Fiege, A. Griewank, and A. Walther. An exploratory line search for piecewise differentiable objective functions based on algorithmic differentiation. In *PAMM*, volume 12, pages 631–632, 2012.

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

[9] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman and Co, 1979.

[10] I. Ginchev and B. Mordukhovich. Directional subdifferentials and optimality conditions. *Positivity*, 16(4):707–737, 2012.

[11] A. Griewank. On stable piecewise linearization and generalized algorithmic differentiation. *Opt. Meth. and Softw.*, 28(6):1139–1178, 2013.

[12] A. Griewank, J.-U. Bernt, M. Randons, and T. Streubel. Solving piecewise linear systems in abs-normal form. *Linear Algebra and its Applications*, 471:500–530, 2013.

[13] A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM, 2008.

[14] A. Griewank, A. Walther, and S. Fiege. Lipschitz piecewise smooth minimization. Technical report, HU Berlin, 2015.

[15] M. Gürbüzbalaban and M.L. Overton. On Nesterov's nonsmooth Chebyshev-Rosenbrock functions. *Nonlinear Anal: Theory, Methods & Appl.*, 75(3):1282–1289, 2012.

[16] W. Hare and J. Nutini. A derivative-free approximate gradient sampling algorithm for finite minimax problems. *Comput. Optim. Appl.*, 56(1):1–38, 2013.

[17] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer, 1993.

[18] C. Lemaréchal. Nonsmooth optimization and descent methods. Technical Report 78,4, IIASA, 1978.

[19] C. Lemaréchal and C. Sagastizábal. Variable metric bundle methods: from conceptual to implementable forms. *Math. Program.*, 76(3):393–410, 1997.

[20] A. Lewis and M. Overton. Nonsmooth optimization via quasi-Newton methods. *Math. Program.*, 141(1-2):135–163, 2013.

[21] M.M. Mäkelä. Multiobjective proximal bundle method for nonconvex nonsmooth optimization: Fortran subroutine MPBNGC 2.0. Technical Report No. B 13/2003, University of Jyväskylä, 2003.

[22] B. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*. Springer, 2006.

[23] Y. Nesterov. Lexicographic differentiation of nonsmooth functions. *Math. Program.*, 104(2-3):669–700, 2005.

[24] S. Scholtes. *Introduction to piecewise differentiable functions*. Springer, 2012.

[25] J. Shen, L. Han, and J.S. Pang. Switching and stability properties of conewise linear systems. *ESAIM: Control, Optimisation and Calculus of Variations*, pages 764–793, 2010.

[26] N.Z. Shor. *Nondifferentiable optimization and polynomial problems.* Kluwer, 1998.

[27] A. Walther and A. Griewank. *Combinatorial Scientific Computing*, chapter Getting Started with ADOL-C, pages 181–202. Chapman-Hall CRC Computational Science, 2012.

[28] P. Wolfe. A method of conjugate subgradients for minimizing nondifferentiable functions. *Mathematical Programming Studies*, 3:145–173, 1975.

[29] G. Yuan, Z. Wei, and Z. Wang. Gradient trust region algorithm with limited memory BFGS update for nonsmooth convex optimization. *Comp. Opt. Appl.*, 54(1):45–64, 2012.