DFG Deutsche Forschungsgemeinschaft
Priority Programme 1962

Beyond the Oracle: Opportunities of Piecewise Differentiation

Andreas Griewank, Andrea Walther

SPP 1962 | Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization

Preprint Number SPP1962-086

received on October 18, 2018

Edited by
SPP1962 at Weierstrass Institute for Applied Analysis and Stochastics (WIAS)
Leibniz Institute in the Forschungsverbund Berlin e.V.
Mohrenstraße 39, 10117 Berlin, Germany
E-Mail: spp1962@wias-berlin.de

World Wide Web: http://spp1962.wias-berlin.de/

Beyond the Oracle: Opportunities of Piecewise Differentiation

Andreas Griewank and Andrea Walther

Abstract For more than thirty years much of the research and development in nonsmooth optimization has been predicated on the assumption that the user provides an oracle that evaluates at any given $\mathbf{x} \in \mathbb{R}^n$ the objective function value $\varphi(\mathbf{x})$ and a generalized gradient $\mathbf{g} \in \partial\varphi(\mathbf{x})$ in the sense of Clarke. We will argue here that, if there is a realistic possibility of computing a vector $\mathbf{g}$ that is guaranteed to be a generalized gradient, then one must know so much about the way $\varphi : \mathbb{R}^n \to \mathbb{R}$ is calculated that more information about the behavior of φ in a neighborhood of the evaluation point can be extracted. Moreover, the latter can be achieved with reasonable effort and in a stable manner so that the *derivative information* provided varies Lipschitz continuously with respect to $\mathbf{x}$. In particular we describe the calculation of directionally active generalized gradients, generalized ε-gradients and the checking of first and second order optimality conditions. All this is based on the abs-linearization of a piecewise smooth objective in abs-normal form.

1 Motivation and Introduction

It is well understood that the convex set $\partial\varphi(\mathbf{x})$ of generalized gradients is highly volatile with respect to variations in $\mathbf{x}$, never mind that it is by definition outer semi-continuous as a set-valued mapping $\partial\varphi : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. Moreover, due to Rademacher's theorem for Lipschitzian functions on a Euclidean space, we can expect that almost everywhere we get a singleton $\partial\varphi(\mathbf{x}) = \{\partial\varphi(\mathbf{x})\}$ so

Andreas Griewank
School of Mathematical and Computational Sciences, Yachay Tech, Urcuquí, Imbabura, Ecuador, e-mail: griewank@yachaytech.edu.ec

Andrea Walther
Institut für Mathematik, Universität Paderborn, Paderborn, Germany e-mail: andrea.walther@uni-paderborn.de

that the users effort to somehow code up a guaranteed generalized gradient will rarely ever pay off during an optimization run. In fact in the paper [30], it is simply assumed that this exceptional event will never happen, so that at the iterates the actually generated $\mathbf{g}$ will always be a classical Fréchet gradient. However, it must be noted that even when this optimistic assumption holds, nonsmooth functions may be poorly approximated by their tangent plane, because there can be a kink nearby about which the local gradient knows nothing. Fortunately, one can generate a local piecewise linear model that reflects such close-by derivative discontinuities, so that whatever algorithm one uses has a chance to deal with the nonsmoothness appropriately. In other words, we suggest to handle (possibly multiple and deflected) kinks at the level of the piecewise linearization.

This chapter is organized as follows. In the second section, we introduce the basic notation and discuss the relation between the Oracle Paradigm and Piecewise Differentiability. In Section 3, we discuss the framework of objective functions in abs-normal form and in Section 4 our approach to generate a local piecewise linear model. In Section 5, we show how one can get information about the gradients that are active in a neighborhood, in particular the gradients and ε-gradients. While these only allow the checking of stationarity and ε-stationarity, we discuss in Section 6 the issue of testing for criticality. In the same section, we introduce briefly an algorithm that actually allows to reach a stationary point. The various concepts discussed in this chapter are illustrated in Section 7 by means of the Crescent example. Section 8 discusses several ways of generating abs-linear approximations of functions including the Euclidean norm and compare their efficiency on a paradigmatic example. Finally, we give a summary and a conclusion in Section 9. Throughout we will consider only the unconstrained case, but most arguments and results carry over to constrained optimization.

2 The Oracle and Piecewise Differentiation

Throughout this chapter we assume that the objective $\varphi : \mathcal{D} \mapsto \mathbb{R}$ is locally Lipschitz on an open domain $\mathcal{D} \subset \mathbb{R}^n$. Moreover, we will use the notation and terminology

$$\begin{aligned}
&\text{Fréchet gradient:} && \nabla\varphi(\mathbf{x}) \equiv \tfrac{\partial\varphi(\mathbf{x})}{\partial\mathbf{x}} : \mathcal{D} \mapsto \mathbb{R}^n \cup \emptyset \\
&\text{Limiting differential:} && \partial^L\varphi(\mathring{\mathbf{x}}) \equiv \overline{\lim}_{\mathbf{x}\to\mathring{\mathbf{x}}} \nabla\varphi(\mathbf{x}) : \mathcal{D} \rightrightarrows \mathbb{R}^n \\
&\text{Clarke differential:} && \partial\varphi(\mathbf{x}) \equiv \mathrm{conv}(\partial^L\varphi(\mathbf{x})) : \mathcal{D} \rightrightarrows \mathbb{R}^n \ .
\end{aligned}$$

where conv denotes the convex hull. The individual elements of the limiting and the Clarke differential will be called limiting gradients and generalized gradients, respectively. The limiting differential is the outer semi-continuous limit of the Fréchet gradient in the Kuratowski-Painlevé sense [35, Section 4.B] so that we have more precisely

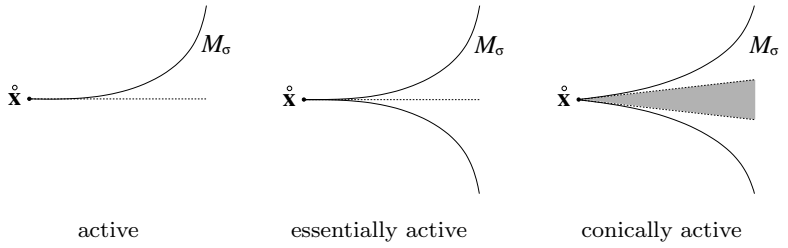

Fig. 1 Different coincidence sets with tangential cones

$$\partial^L \varphi(\mathbf{x}) \equiv \left\{ \lim_{i\to\infty} \nabla\varphi(\mathbf{x}_i) : \mathbf{x}_i \to \mathbf{x}, \mathcal{D} \ni \mathbf{x}_i \notin \mathcal{S} \right\} .$$

Here $\mathcal{S}$ denotes the set of exceptional points, where $\varphi(\mathbf{x})$ is not Fréchet differentiable. In the literature the limiting differential is often called the Bouligand differential or derivative.

Definition 1 (Oracle Paradigm). The locally Lipschitz continuous function $\varphi : \mathbb{R}^n \mapsto \mathbb{R}$ is said to satisfy the Oracle Paradigm if at any $\mathbf{x} \in \mathbb{R}^n$ not only the function value $\varphi(\mathbf{x})$ but also at least one generalized gradient $\mathbf{g} \in \partial\varphi(\mathbf{x})$ can be made available to the optimization algorithm.

At first the task required by the oracle paradigm does not appear that hard in the piecewise differentiable case.

Definition 2 (Piecewise Differentiability). The locally Lipschitz continuous function $\varphi : \mathcal{D} \subset \mathbb{R}^n \mapsto \mathbb{R}$ is said to be $d > 0$ times piecewise differentiable if at any $\mathbf{x} \in \mathcal{D}$ there exists a selection function $\varphi_\sigma(\mathbf{x}) \in C^d(\mathcal{D})$ such that $\varphi(\mathbf{x}) = \varphi_\sigma(\mathbf{x})$. Here σ belongs to some finite index set $\mathcal{E}$ labeling the selection functions φ_σ.

In the literature the elements of $\mathcal{E}$ are usually chosen as natural numbers, but we will give ourselves a little more freedom and later define them as tuples of a certain kind.

Definition 3 (Active Selections). The selection function φ_σ is said to be *active* at $\mathring{\mathbf{x}} \in \mathcal{D}$ if the *coincidence set* $M_\sigma = \{\mathbf{x} \in \mathcal{D} : \varphi(\mathbf{x}) = \varphi_\sigma(\mathbf{x})\}$ contains $\mathring{\mathbf{x}}$. Moreover φ_σ is called *essentially active* if $\mathring{\mathbf{x}}$ belongs to the closure of the interior of M_σ. Finally, it is called *conically active* if the tangent cone of M_σ at $\mathring{\mathbf{x}}$ has a nonempty interior, The index sets of the correspondingly active selection function indexes will be denoted by the chain $\mathcal{E} \supset \mathcal{E}_a(\mathring{\mathbf{x}}) \supset \mathcal{E}_e(\mathring{\mathbf{x}}) \supset \mathcal{E}_c(\mathring{\mathbf{x}})$.

The inclusion relations at the end of the last definition are easily verified. They become intuitively clear if one looks at the drawings in Figure 1.

Lemma 1 (Scholtes [36]). *In the piecewise smooth case the limiting differential is the span of the essentially active gradients, i.e.,*

$$\partial^L \varphi(\mathring{\mathbf{x}}) \;=\; \bigcup_{\sigma \in \mathcal{E}_e(\mathbf{x})} \{\nabla \varphi_\sigma(\mathring{\mathbf{x}})\}$$

whose convex hull is of course the Clarke differential.

To realize the oracle here one would have to find a signature that is not only active but essentially active, which does not seem quite so simple. For our class it turns out to be easier to get the under and overestimations

$$\emptyset \neq \partial^K \varphi(\mathring{\mathbf{x}}) \;\equiv \bigcup_{\sigma \in \mathcal{E}_c(\mathbf{x})} \{\nabla \varphi_\sigma(\mathring{\mathbf{x}})\} \;\subset\; \partial^L \varphi(\mathring{\mathbf{x}}) \;\subset \bigcup_{\sigma \in \mathcal{E}_a(\mathbf{x})} \{\nabla \varphi_\sigma(\mathring{\mathbf{x}})\} \;.$$

The first set on the left will be called the conic differential, as it contains only gradients $\nabla \varphi_\sigma(\mathring{\mathbf{x}})$ of selection functions that are conically active at $\mathring{\mathbf{x}}$. As we will see $\partial^K \varphi(\mathring{\mathbf{x}})$ is never empty and we will be able to compute one or even all of its finitely many elements for objectives in abs-normal form. It might be reasonably claimed that only the conically active gradients are relevant for optimizing φ in the vicinity of $\mathring{\mathbf{x}}$.

The last set on the right can be a gross-overestimation which might come about if one applies the generalized differentiation rules forward in a naive way. To avoid this overestimation one has to detect all selection functions that are active but not essentially active, a rather daunting task for a set of nonlinear $\varphi_\sigma(\mathbf{x})$ as the coincidence sets M_σ may be very complicated even if all $\varphi_\sigma(\mathbf{x})$ are assumed to be polynomial. Then realizing the oracle paradigm must be considered rather difficult.

A challenging question is how we can evaluate the multifunctions

$$\mathcal{E}_a : \mathbf{x} \in \mathcal{D} \rightrightarrows \mathcal{E}, \quad \mathcal{E}_e : \mathbf{x} \in \mathcal{D} \rightrightarrows \mathcal{E}, \quad \mathcal{E}_c : \mathbf{x} \in \mathcal{D} \rightrightarrows \mathcal{E} \;.$$

The first one appears easy except that testing equality in floating point arithmetic is always a little dicey. A popular format used in some software packages is that the coincidence sets are defined by $|\mathcal{E}|$ different systems of linear (or nonlinear) inequalities like

$$\sigma \in \mathcal{E}_a(\mathbf{x}) \iff A_\sigma \mathbf{x} \leq \mathbf{b}_\sigma \quad \text{for} \quad A_\sigma \in \mathbb{R}^{m_\sigma \times n},\; \mathbf{b}_\sigma \in \mathbb{R}^{m_\sigma} \;.$$

The difficulty with this representation is that it suffers from three related drawbacks.

- Likely exponential size of data structure.
- Redundancy because pieces must fit together.
- Numerical perturbations or typos destroy consistency.

Of course, everything is much worse in the nonlinear case. We view the representation by pieces as one of the main reasons why piecewise linear or smooth functions have not been used very much in scientific computing. Let us look at a small example.

The half pipe example

Firstly we consider the definition by pieces

$$\varphi : \mathbb{R}^2 \mapsto \mathbb{R}, \qquad \varphi(x_1, x_2) = \begin{cases} \varphi_{-1,1}(x_1, x_2) = x_2^2 & \text{if } x_1 \leq 0 \\ \varphi_{1,-1}(x_1, x_2) = x_2^2 - x_1 & \text{if } 0 \leq x_1 \leq x_2^2 \\ \varphi_{1,1}(x_1, x_2) = 0 & \text{if } 0 \leq x_2^2 \leq x_1 \end{cases} .$$

The graph of this function is given by Figure 2.

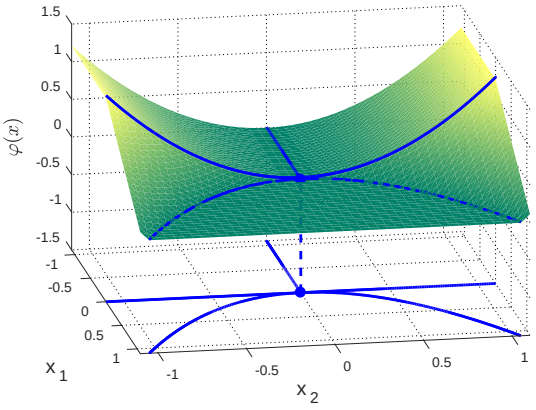

Fig. 2 Half pipe function with a priori five smooth pieces

The double indexing of the functions will become clearer later. The corresponding coincidence sets $\mathcal{S}_{-1,1}$, $\mathcal{S}_{1,-1}$, $\mathcal{S}_{1,1}$ are all essentially active at the origin, but the double cusp shaped one in the middle is not conically active. The corresponding gradients are

$$\nabla\varphi_{-1,1}(0,0) = (0,0) = \nabla\varphi_{1,1}(0,0) \quad \text{and} \quad \nabla\varphi_{1,1}(0,0) = (-1,0) .$$

Hence we get the differentials

$$\{\nabla\varphi(0,0)\} \;=\; \{(0,0)\} \;=\; \partial^K\varphi(0,0) \subsetneq \partial^L\varphi(0,0) \;=\; \{(0,0),(-1,0)\} .$$

As we see the function is actually differentiable at the origin, which is reflected in the conic differential being a singleton containing only the Fréchet gradient $(0,0)$. In contrast the limiting differential picks up the gradient $(-1,0)$ from the double cusp $\mathcal{S}_{1,-1}$, which is not conically active. Of course the Clarke differential as the convex hull of the limiting differential is simply the line segment $\partial\varphi(0,0) \;=\; \{(\alpha,0) : \alpha \in [-1,0]\}$. To highlight the role of nonessentially active selection functions let us introduce the function $\varphi_{0,1}(x_1,x_2) = x_2^2 + x_1$ with the coincidence set $\mathcal{S}_{0,1} = \{0\} \times \mathbb{R}$. This selection function is active at

the origin but its gradient $\nabla\varphi_{0,1}(0,0) = (0,1)$ does and may not belong to the Clarke differential $[-1,0] \times 0$. Hence it would be a failure for the oracle to return the gradient $(0,1)$ of an active selection function as a generalized let alone limiting gradient.

Generally this difficulty arises because generalized differentiation rules are mostly only inclusions. Only the operations that are guaranteed to maintain convexity with respect to $\mathbf{x}$, namely conic combinations and pointwise maximization, i.e.,

$$v(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i u_i(\mathbf{x}) \quad \text{and} \quad v(\mathbf{x}) = \max_{i=1\ldots m} \{\alpha_i u_i(\mathbf{x})\} \quad \text{with}$$
$$\alpha_i \geq 0 \quad \text{for } i = 1, \ldots, m$$

also propagate the corresponding generalized gradients as identities, see [7], such that

$$\partial v(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \partial u_i(\mathbf{x}) \quad \text{and} \quad \partial v(\mathbf{x}) = \operatorname{conv}\{\alpha_i \partial u_i(\mathbf{x}) : u_i(\mathbf{x}) = v(\mathbf{x})\} \ ,$$

respectively. All other elementary operations, in particular subtraction and multiplication propagate generalized derivatives only as inclusions, i.e., we have

$$\partial(u - w) \subset \operatorname{conv}\{\partial u - \partial w\} \qquad \text{and} \qquad \partial(u \cdot w) \ \subset \ \operatorname{conv}\{w\,\partial u + u\,\partial w\} \ ,$$

where we have left off the argument $\mathbf{x}$ for notational simplicity. The same is true for the absolute value function $v = \operatorname{abs}(x)$ so that one can never be sure to obtain true generalized gradients when propagating such vectors forward through a chain of operations.

At a higher level, compositions of vector functions propagate generalized derivatives by the chain rule also as an inclusion, with identity holding only when one of the factors involved is smooth or at least subdifferentially regular [3, Def. 3.5]. The verification of this frequently assumed property was shown to be co-NP complete in [39] on the class of piecewise smooth functions $C^d_{\text{abs}}(\overline{\mathcal{D}})$ defined below. It actually amounts to local convexity of the piecewise linearization or equivalently the directional derivative $\varphi'(\mathbf{x};\cdot)$ and is thus a rather strong and complex assumption. Finally, we note that for approximating generalized gradients by divided differences, see [3, Chap. 6], one has to rely on this convex set being finitely generated and thus polyhedral, which comes pretty close to assuming abs-normality as defined in the next section. We will see, that assumption allows in fact the exact calculation of the conical differential $\partial^K\varphi(\mathbf{x})$ which is an always nonempty subset of the the limiting differential $\partial^L\varphi(\mathbf{x})$.

3 Abs-normal objectives

For the half pipe example, one may consider several formulations. Firstly, what one might consider the 'Original' formulation in terms of max

$$\varphi(x_1, x_2) \ = \max(x_2^2 - \max(x_1, 0), 0) \ .$$

Here the Lipschitz continuity is immediately apparent. Alternatively, rewriting max in terms of abs we get after some mechanical manipulations

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

Now we have a formulation where all nonsmoothness is cast in terms of the absolute value function, which occurs at the two arguments x_1 and $x_2^2 - \frac{1}{2}(x_1 + |x_1|)$. Wherever these quantities change their sign we will have a kink in the function value. Therefore we name them switching variables and define them as

$$(z_1, z_2) = F(x_1, x_2, z_1) = \left(x_1, x_2^2 - \tfrac{1}{2}(x_1 + |z_1|) \right) \ . \tag{1}$$

Substituting them into the original expression we get

$$\varphi(x_1, x_2) = f(\mathbf{x}, |\mathbf{z}|) = \tfrac{1}{2} \left(x_2^2 - \tfrac{1}{2}(x_1 + |z_1|) + |z_2| \right) \ . \tag{2}$$

Now we have two equations (1) and (2) that define the half pipe function in a nonredundant and stable way. Any of the coefficients, which are mostly 1 can be perturbed with the resulting function still being well defined and Lipschitz continuous. Moreover, we can label the various smooth function pieces by the vector $\sigma = (\mathbf{sgn}(z_1), \mathbf{sgn}(z_2)) \in \{-1, 0, 1\}^2$, which is consistent with the labeling we used in the previous section.

More generally, we will consider the class of objective functions that are defined as compositions of smooth elemental functions and the absolute value function $\mathrm{abs}(x) = |x|$. Hence they may also include $\max(x, y), \min(x, y)$, and the positive part function $\mathrm{pos}(x) \equiv \max(x, 0)$, which can all be easily cast in terms of an absolute value. By successively numbering all arguments of absolute value evaluations as *switching variables* z_i for $i = 1 \ldots s$, we obtain a piecewise smooth representation of $y = \varphi(\mathbf{x})$ in the abs-normal form

$$\mathbf{z} \ = \ F(\mathbf{x}, |\mathbf{z}|) \tag{3}$$

$$y \ = \ f(\mathbf{x}, |\mathbf{z}|) \ , \tag{4}$$

where for $\mathcal{D} \subset \mathbb{R}^n$ open, $F : \overline{\mathcal{D}} \times \overline{\mathbb{R}^s_+} \mapsto \mathbb{R}^s$ and $f : \overline{\mathcal{D}} \times \overline{\mathbb{R}^s_+} \mapsto \mathbb{R}$ with $\overline{\mathcal{D}} \times \overline{\mathbb{R}^s_+} \subset \mathbb{R}^{n+s}$. Here, z_j can only influence z_i if $j < i$ so that when interpreting F as a function of $|\mathbf{z}|$, its Jacobian with respect to $|\mathbf{z}|$ is strictly lower triangular. Consequently, we can evaluate for any $\mathbf{x}$ the unique, piecewise smooth value $\mathbf{z}(\mathbf{x})$. In other words, we state the calculation of all switching

variables as equality constraints and handle the vector of the absolute values of the switching variables as extra argument of the then smooth target function f. Sometimes, we write

$$\varphi(\mathbf{x}) \;\equiv\; f(\mathbf{x}, |\mathbf{z}(\mathbf{x})|)$$

to denote the objective directly in terms of the argument vector $\mathbf{x}$ only. In this chapter, we are mostly interested in the case where the nonlinear elementals are all once or twice continuously differentiable. The resulting function class was first considered in [12] and is specified as follows:

Definition 4. For any $d \in \mathbb{N}$ and $\mathcal{D} \subset \mathbb{R}^n$, the set of functions $\varphi : \overline{\mathcal{D}} \mapsto \mathbb{R}$ defined by an abs-normal form (3)-(4) with $f, F \in C^d(\overline{\mathcal{D}} \times \overline{\mathbb{R}^s_+})$ is denoted by $C^d_{\rm abs}(\overline{\mathcal{D}})$.

Recall that $C^d(\overline{\Omega})$ is the set of functions that possess continuous d-th derivatives in the open set Ω that can be continuously extended to the boundary $\partial\Omega = \overline{\Omega}\backslash\Omega$. In the usual case, where F and f are themselves compositions of smooth elemental functions φ_i these are assumed to be $C^d(\overline{\mathcal{D}}_i)$ functions on their respective domains $\mathcal{D}_i$ reachable from $\mathbf{x} \in \mathcal{D}$. The combinatorial structure of the nonsmooth function φ can be described by the signature vector and matrix

$$\sigma = \sigma(\mathbf{x}) \equiv \mathbf{sgn}(\mathbf{z}(\mathbf{x})) \in \{-1, 0, +1\}^s \;\text{ and }\; \Sigma \equiv \Sigma(\mathbf{x}) = \mathrm{diag}(\sigma(\mathbf{x})) \in \mathbb{R}^{s\times s}\,.$$

For fixed σ and the corresponding Σ we can locally solve Eq. (3) using $|\mathbf{z}(\mathbf{x})| = \Sigma\mathbf{z}(\mathbf{x})$ for $\mathbf{z}(\mathbf{x})$ and thus have the implicitly defined selection function

$$\varphi_\sigma(\mathbf{x}) \;\equiv\; f(\mathbf{x}, \Sigma\mathbf{z}(\mathbf{x})) \quad \text{s.t.} \quad \mathbf{z}(\mathbf{x}) = F(\mathbf{x}, \Sigma\mathbf{z}(\mathbf{x}))\,. \tag{5}$$

Again due to the assumed triangularity, the system of equations is locally solvable for $\mathbf{z}(\mathbf{x})$ by the implicit function theorem. Hence, the functions in $C^d_{\rm abs}(\overline{\mathcal{D}})$ are certainly piecewise smooth as defined in Definition 2. In our scenario $\mathcal{E}$ is a subset of $\{-1, 0, 1\}^s$ by definition of σ. Generally in the literature, it remains a little mysterious how a suitable index σ is chosen as a function of $\mathbf{x}$ such that the resulting function is continuous. In our function model the determination of the $\sigma(\mathbf{x})$ is intertwined with the computation of the numerical values.

Related piecewise linear functions

The class $C^d_{\rm abs}(\overline{\mathcal{D}})$ covers many piecewise smooth functions but certainly not all. For example, on a given triangulation of the plane or space, somebody may have spliced together different local models such that they fit continuously across the triangle edges or tetrahedron faces. In this situation it seems impossible to deduce from the properties of the function within one triangle

or tetrahedron anything about what is happening in the neighboring triangles or tetrahedrons, let alone further afield.

In contrast, in some sense the functions belonging to $C^d_{\rm abs}(\overline{\mathcal{D}})$ allow extrapolation from one polyhedron to its neighbors. We even harbor the hope that there might be reasonably efficient methods to globally optimize piecewise linear functions using their abs-linear form. That representation always exists but may be not so easy to construct.

On the other hand, we must admit that in the spliced situation the oracle paradigm appears quite natural with each limiting differential being just the gradient of the adjacent patches, i.e., one of the selection functions. However, except on triangular or quadrilateral grids it may again not be easy to decide which selection function is essentially active as defined in Definition 3. Of course any statement of stationarity or optimality will only apply to the patch and nothing can be said about the behavior of the piecewise smooth function in an open neighborhood, no matter how small.

Another class of possibly even linear piecewise smooth functions that does not fit within the $C^d_{\rm abs}(\overline{\mathcal{D}})$ framework are solution operators like

$$\varphi(\mathbf{x}) \;=\; \max\{\tfrac{1}{2}\mathbf{y}^T Q\mathbf{y} + \mathbf{c}^T\mathbf{y} : A\mathbf{y} \le B\mathbf{x} + \mathbf{c}\} \quad \text{with} \quad Q = Q^T \succ 0\,.$$

Here we may use a finite solver to compute the mathematically well defined piecewise linear function $\varphi(\mathbf{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ but the number of steps may vary depending on the argument $\mathbf{x}$. Moreover, there may be degeneracies whose numerical resolution requires if-statements and other program branches. As shown in [4], for implicitly defined functions like $G(\mathbf{x}, \mathbf{y}(\mathbf{x})) = 0$ with G in abs-normal form one, can compute its abs-linear approximation $\Delta\mathbf{y}(\mathring{\mathbf{x}}, \Delta\mathbf{x})$ by a generalized version of the implicit function theorem. Albeit with some nontrivial effort, this could be integrated into the extended AD software, which has not been done.

In the penultimate section of the paper we will consider the extension of $C^d_{\rm abs}(\overline{\mathcal{D}})$ to its superset $C^d_{\rm euc}(\overline{\mathcal{D}})$, which consists of all functions that can be evaluated as compositions of C^d elementals and the Euclidean norm $\|\cdot\| = \|\cdot\|_2$. These Lipschitz continuous functions are no longer piecewise smooth, but one can still construct abs-linear approximations that appear to be useful for optimization purposes.

It has recently been observed [21] that minimizing a function in abs-normal form is equivalent to solving the equality constrained MPEC

$$\begin{aligned} &\min \varphi = f(\mathbf{x}, \mathbf{u} + \mathbf{v}) \\ &\text{s.t. } 0 \le \mathbf{u} \perp \mathbf{v} \ge 0 \\ &\mathbf{u} - \mathbf{v} = F(\mathbf{x}, \mathbf{u} + \mathbf{v})\,. \end{aligned}$$

Here, MPEC means Mathematical Programming with Equilibrium Constraints [31]. Notice that in general, i.e., without the triangularity of F with respect to $|\mathbf{z}| = \mathbf{u} + \mathbf{v}$, the MPEC may be quite hard to solve, if only because

one cannot easily compute a feasible $\mathbf{z} = \mathbf{u} - \mathbf{v}$ for any given $\mathbf{x}$. Nevertheless, it was shown in [21] that in the triangular case the constraint qualification MPEC-LICQ is equivalent to the Linear Independent Kink Qualification introduced in [17] for abs-normal objectives.

4 The abs-linear approximation

All abs-normal objectives are strongly semi-smooth as defined in [8, Chap. 7] and thus their generalized gradients satisfy for fixed $\mathring{\mathbf{x}}$ the backward approximation property [22]

$$\varphi(\mathbf{x}) - \varphi(\mathring{\mathbf{x}}) - \mathbf{g}^T(\mathbf{x} - \mathring{\mathbf{x}}) \;=\; \mathcal{O}(\|\mathbf{x} - \mathring{\mathbf{x}}\|^2) \quad \text{for all} \quad \mathbf{g} \in \partial\varphi(\mathbf{x}) \,.$$

While the vector version of this relation forms the basis of the semi-smooth Newton method it is not clear how it can be exploited for the purposes of unconstrained local optimization. Instead we aim for a generalization of the classical first order Taylor expansion, which is in some sense forward, from the reference point $\mathring{\mathbf{x}}$ with the corresponding $\mathring{\mathbf{z}} = \mathbf{z}(\mathring{\mathbf{x}})$ and $\mathring{\mathbf{y}} = y(\mathring{\mathbf{x}})$ to a trail point $\mathbf{x} \approx \mathring{\mathbf{x}}$ and the corresponding $\mathbf{z} = \mathbf{z}(\mathbf{x})$ and $y = y(\mathbf{x})$. From Equations (3) and (4), one obtains the smooth Taylor expansion

$$\begin{bmatrix} \mathbf{z} - \mathring{\mathbf{z}} \\ y - \mathring{\mathbf{y}} \end{bmatrix} \;=\; \begin{bmatrix} Z & L \\ \mathbf{a}^T & \mathbf{b}^T \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mathring{\mathbf{x}} \\ |\mathbf{z}| - |\mathring{\mathbf{z}}| \end{bmatrix} \;+\; \mathcal{O}\left(\begin{bmatrix} \|\mathbf{x} - \mathring{\mathbf{x}}\|^2 \\ \|\mathbf{z} - \mathring{\mathbf{z}}\|^2 \end{bmatrix} \right)$$

as abs-linear approximation of $y = \varphi(\mathbf{x})$. Here the matrices

$$L \equiv \frac{\partial F(\mathbf{x}, |\mathbf{z}|)}{\partial |\mathbf{z}|} \in \mathbb{R}^{s \times s}, \quad Z \equiv \frac{\partial}{\partial \mathbf{x}} F(\mathbf{x}, |\mathbf{z}|) \in \mathbb{R}^{s \times n} \tag{6}$$

and the vectors

$$\mathbf{a} = \frac{\partial}{\partial x} f(\mathbf{x}, |\mathbf{z}|) \in \mathbb{R}^n, \quad \mathbf{b} = \frac{\partial}{\partial |\mathbf{z}|} f(\mathbf{x}, |\mathbf{z}|) \in \mathbb{R}^s \tag{7}$$

are evaluated at the reference point $(\mathring{\mathbf{x}}, \mathring{\mathbf{z}})$. Due to the triangularity of F and thus L one can easily check by induction that $\|\mathbf{z} - \mathring{\mathbf{z}}\| = \mathcal{O}(\|\mathbf{x} - \mathring{\mathbf{x}}\|)$. Hence we obtain with $\Delta\mathbf{x} \equiv \mathbf{x} - \mathring{\mathbf{x}}$ and

$$\tilde{\mathbf{z}} \equiv \mathbf{c} + Z\Delta\mathbf{x} + L|\tilde{\mathbf{z}}| \equiv (\mathring{\mathbf{z}} - L|\mathring{\mathbf{z}}|) + Z\Delta\mathbf{x} + L|\tilde{\mathbf{z}}| \tag{8}$$

the incremental approximation

$$\Delta\varphi(\mathring{\mathbf{x}}; \Delta\mathbf{x}) \equiv \mathbf{a}^T\Delta\mathbf{x} + \mathbf{b}^T(|\tilde{\mathbf{z}}| - |\mathring{\mathbf{z}}|) \;=\; y - \mathring{\mathbf{y}} + \mathcal{O}(\|\Delta\mathbf{x}\|^2) \tag{9}$$

or equivalently

$$\varphi(\mathring{\mathbf{x}} + \Delta\mathbf{x}) - \varphi(\mathring{\mathbf{x}}) \;=\; \Delta\varphi(\mathring{\mathbf{x}}; \Delta\mathbf{x}) + \mathcal{O}(\|\Delta\mathbf{x}\|^2) \;. \tag{10}$$

In other words, we have a generalized Taylor approximation with uniform error term $\mathcal{O}(\|\Delta\mathbf{x}\|^2)$, which is in contrast to directional differentiation completely independent of the direction $\Delta\mathbf{x}/\|\Delta\mathbf{x}\|$. This is possible because $\Delta\varphi(\mathring{\mathbf{x}}; \Delta\mathbf{x})$ is with respect to $\Delta\mathbf{x}$ piecewise linear but in contrast to the directional derivative $\varphi'(\mathring{\mathbf{x}}; \Delta\mathbf{x})$ not homogeneous. That means, it "knows" about nearby kinks, which is exactly the kind of information that any kind of strictly local generalized differentiation can not pick up. In previous papers [13], [15] we have derived the relation (10) by induction on the intermediate quantities occurring in the evaluation procedure of the overall $\varphi(\mathbf{x})$. This approach then directly yields the partial elemental derivatives, from which the "global" matrices and vectors $[L, Z, \mathbf{a}, \mathbf{b}, \mathbf{c}]$ can be accumulated by suitable variants of the chain rule.

For example let us consider the half pipe example in the abs-normal form (1) and (2). Then, we get by differentiation at any $\mathring{\mathbf{x}}$

$$L = \begin{bmatrix} 0 & 0 \\ -\frac{1}{2} & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} -0.25 \\ 0.5 \end{bmatrix}$$

and in dependence on the reference point $\mathring{\mathbf{x}} = (\mathring{x}_1, \mathring{x}_2)$ with $\mathring{\mathbf{z}} = \mathbf{z}(\mathring{\mathbf{x}})$

$$Z = \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & 2\,\mathring{x}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{a} \;=\; \begin{bmatrix} -0.25 \\ \mathring{x}_2 \end{bmatrix} \;.$$

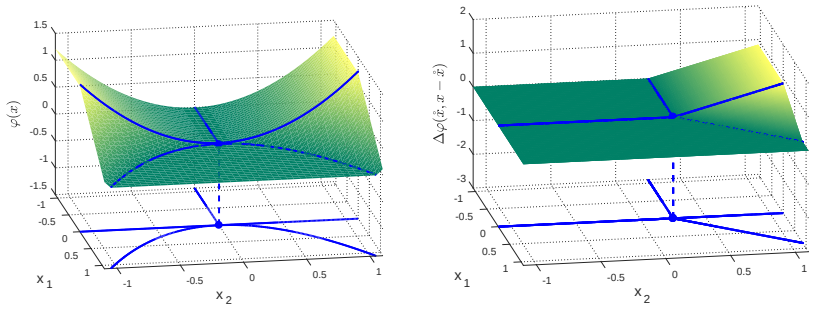

Fig. 3 The half pipe function and its abs-linearisation at $\mathring{\mathbf{x}} = (-1, 1)$

The original function and the resulting abs-linearization at $\mathring{\mathbf{x}} = (-1, 1)$ are illustrated in Figure 3. At $\mathring{\mathbf{x}}$ we have $\mathring{\mathbf{z}} = (-1, 1)$ and thus $\mathring{\sigma} = (-1, 1)$ which means that the function is locally completely smooth at $\mathring{\mathbf{x}}$ but the abs-linearization still has an idea where there are kinks. Notice that Z has the determinant $2x_2$ which means that at the origin where both switching

variables are active the Linear-Independent-Kink Qualification, as introduced in [17] is not satisfied.

Evaluating $\Delta\varphi(\mathring{\mathbf{x}};\Delta\mathbf{x})$ via Equation (8) and (9) is quite cheap, provided the matrices and vectors $(Z, L, \mathbf{a}, \mathbf{b}, \mathbf{c})$, which constitute the abs-linear approximation are known. They are all first derivatives of smooth, composite functions, so they can be obtained by algorithmic or automatic differentiation. In fact the well known AD tools ADOL-C [38], CPP-AD [5] and Tapenade [20] have been extended to generate abs-linear approximations. Of course, just like in the smooth case, where the evaluation of the full Jacobian can be avoided through iterative methods that are based exclusively on tangents in the sense of Jacobian×vector products and co-tangents or adjoints in the sense of row-vector×Jacobian products, such a matrix free approach can also be pursued for the abs-linear approximation. However, for notational simplicity we will assume in this chapter that the matrices Z and L are completely accumulated. As defined via Equations (8) and (9) for fixed $\mathring{\mathbf{x}}$ the function $\Delta\varphi(\mathring{\mathbf{x}};\Delta\mathbf{x})$ is just an abs-normal function, where all operations other than the absolute value are linear or more precisely affine.

5 Checking gradient activity

Now the question arises what information about $\varphi(\mathbf{x})$ near the reference point $\mathring{\mathbf{x}}$ we can gain from the analysis of its abs-linear approximation $\Delta\varphi(\mathring{\mathbf{x}};\Delta\mathbf{x})$ near $\Delta\mathbf{x} = 0$. For notational simplicity we set $\mathring{\mathbf{x}} = 0$, replace $\Delta\mathbf{x}$ by $\mathbf{x}$ and $\tilde{\mathbf{z}}$ by $\mathbf{z}$ as well as ignoring constant shifts in the objective function. Then we have simply the abs-linear minimization problem

$$\min \Delta y(\mathbf{x}) \;\equiv\; \mathbf{a}^T\mathbf{x} + \mathbf{b}^T|\mathbf{z}| \qquad \text{s.t.} \qquad \mathbf{z} \equiv \mathbf{c} + Z\mathbf{x} + L|\mathbf{z}| \;. \tag{11}$$

Due to the strict lower triangularity of L there is a unique piecewise linear $\mathbf{z} = \mathbf{z}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$, which is a special case of the piecewise smooth $\mathbf{z}(\mathbf{x})$ considered before for the abs-normal φ itself. However, on the abs-linear level we have a much better chance of dealing with the nonsmoothness represented by the kinks explicitly. Then the full domain $\mathbb{R}^n$ is decomposed into polyhedra, which can be identified by the signature vector and matrix

$$\sigma = \sigma(\mathbf{x}) \equiv \mathbf{sgn}(\mathbf{z}(\mathbf{x})) \in \{-1, 0, +1\}^s \quad \text{and} \quad \Sigma \equiv \Sigma(\mathbf{x}) = \mathrm{diag}(\sigma(\mathbf{x})) \in \mathbb{R}^{s\times s}$$

now as a function of the piecewise linear $\mathbf{z}(\mathbf{x})$. The inverse images

$$P_\sigma \;\equiv\; \{\mathbf{x} \in \mathbb{R}^n : \sigma(\mathbf{x}) = \sigma\} \tag{12}$$

are pairwise disjoint, relatively open polyhedra. Using the partial order of the signatures given by

$$\tilde{\sigma} \prec \sigma \quad \Longleftrightarrow \quad \tilde{\sigma}_i \sigma_i \leq \sigma_i^2 \quad \text{for} \quad i = 1 \ldots s \;,$$

we can define the essential closures

$$\bar{P}_\sigma \;\equiv\; \{\mathbf{x} \in \mathbb{R}^n : \sigma(\mathbf{x}) \prec \sigma\} \;,$$

which are no longer disjoint and whose inclusion ordering corresponds exactly to the partial ordering $\prec$ of the signatures such that

$$\bar{P}_\sigma \;\subset\; \bar{P}_{\tilde{\sigma}} \;\Longleftrightarrow\; \sigma \prec \tilde{\sigma} \;.$$

Hence, we see that $\mathring{\mathbf{x}} = 0$ with $\mathring{\sigma} = \sigma(\mathring{\mathbf{x}})$ belongs exactly to the essential closures $\bar{P}_\sigma$ for which $\sigma \succ \mathring{\sigma}$. Consequently, we find for some open ball $B(\mathring{\mathbf{x}}; \rho)$

$$B(\mathring{\mathbf{x}}; \rho) = \left\{ \bigcup_{\sigma \succ \mathring{\sigma}} P_\sigma \right\} \cap B(\mathring{\mathbf{x}}; \rho) \;.$$

Here, the σ on the right hand side can be restricted to be definite, i.e., only have nonzero components $\sigma_i = \pm 1$, which will be denoted by $\sigma \not\ni 0$. Within each P_σ, we have $|\mathbf{z}| = \Sigma \mathbf{z}$ so that one can solve the equality constraint on the right hand side of Equation (11) for $\mathbf{z}$ to obtain the affine function

$$\mathbf{z}(\mathbf{x}) = (I - L\Sigma)^{-1}(\mathbf{c} + Z\mathbf{x}) \quad \text{for} \quad \mathbf{x} \in \bar{P}_\sigma \;. \tag{13}$$

Note that due to the strict lower triangularity of L the unit lower triangular matrix $(I - L\Sigma)^{-1}$ is for any σ well defined and its elements are polynomial in the entries of L. For definite signatures $\sigma \not\ni 0$ the elements $x \in \bar{P}_\sigma$ are exactly characterized as solutions of the system of inequalities

$$\Sigma(I - L\Sigma)^{-1}(\mathbf{c} + Z\mathbf{x}) \;=\; (\Sigma - L)^{-1}(\mathbf{c} + Z\mathbf{x}) \;\geq\; 0 \;.$$

If there is an $\mathbf{x} \in \bar{P}_\sigma$ with definite signature $\sigma(\mathbf{x}) \not\ni 0$ then the polyhedron P_σ has a nonempty interior. The converse needs not be true in the presence of degeneracy. From duality theory it is known that either, $\bar{P}_\sigma$ has a nonempty interior, in which case we call it full-dimensional, or the rows of $(\Sigma - L)^{-1} Z$ have a vanishing convex combination such that

$$\lambda^T (\Sigma - L)^{-1} Z = 0 \quad \text{with} \quad 0 \leq \lambda \neq 0 \;.$$

Obviously this can be checked by standard Linear Optimization techniques. One can also check whether $\dim(\bar{P}_\sigma) = n$ in which case we have the gradient

$$\mathbf{g}_\sigma \;=\; \mathbf{a}^T + \mathbf{b}^T \Sigma (I - L\Sigma)^{-1} Z \;=\; \mathbf{a}^T + \mathbf{b}^T (\Sigma - L)^{-1} Z \;, \tag{14}$$

where the last equality relies on definiteness, i.e., $\sigma \not\ni 0$, so that $\det(\Sigma) \pm 1$.

Hence we obtain for the abs-linear approximation the nonempty set of limiting gradients

$$\partial^L_{\Delta \mathbf{x}} \Delta\varphi(\mathring{\mathbf{x}}; \Delta\mathbf{x})\big|_{\Delta\mathbf{x}=0} \;=\; \bigcup_{0 \not\succeq \sigma \succ \mathring{\sigma}} \{\mathbf{a}^T + \mathbf{b}^T \Sigma (I - L\Sigma)^{-1} Z\} \; . \tag{15}$$

Conic activity

Now, let φ be again a general nonlinear $C^d_{\rm abs}$ function. It was shown in [13] and [27] that

$$\emptyset \;\neq\; \partial^K_{\mathbf{x}} \varphi(\mathbf{x})\big|_{\mathbf{x}=\mathring{\mathbf{x}}} \equiv \partial^L_{\Delta \mathbf{x}} \Delta\varphi(\mathring{\mathbf{x}}, \Delta\mathbf{x})\big|_{\Delta\mathbf{x}=0} \;\subset\; \partial^L_{\mathbf{x}} \varphi(\mathbf{x})\big|_{\mathbf{x}=\mathring{\mathbf{x}}} \quad ,$$

which immediately implies the corresponding inclusion for the Clarke differential as convex hull of the limiting differential. The limiting gradients $\mathbf{g}_\sigma \in \partial^K \varphi(\mathring{\mathbf{x}})$ are *conic* gradients of φ as defined in Definition 3. Then we have in fact $\mathbf{g}_\sigma = \partial\varphi_\sigma(\mathring{\mathbf{x}})$ for some $\sigma \in \mathcal{E}_c(\mathring{\mathbf{x}})$. The limiting gradient $\partial^L \varphi(\mathring{\mathbf{x}})$ may contain other gradients $\partial\varphi_{\tilde\sigma}(\mathring{\mathbf{x}})$ of selection functions $\varphi_{\tilde\sigma}$ that are essentially active so that $\tilde\sigma \in \mathcal{E}_e(\mathring{\mathbf{x}}) \setminus \mathcal{E}_c(\mathring{\mathbf{x}})$. If φ happens to be differentiable, but not necessarily strictly differentiable at $\mathring{\mathbf{x}}$ we have simply $\partial^K \varphi(\mathbf{x}) = \{\partial\varphi(\mathbf{x})\}$, which must be the case at almost all points in $\mathbb{R}^n$ by Rademacher's theorem. This applies also to functions like $\varphi(x) = |1 - \sin^2(x) - \cos^2(x)|$ where a conventional chain rule oriented "Are we differentiable at this point?" test, whose use is for example suggested in [6] would naturally always respond "no".

Back to the oracle

Obviously, the observations above also hold for generalized Jacobians of vector-valued functions, so abs-linearization also provides a practical procedure for implementing the semi-smooth Newton method [23, 28]. Then, only one element of $\partial^L \varphi(\mathring{\mathbf{x}}) \subset \partial\varphi(\mathring{\mathbf{x}})$ is required. The same holds true for the subgradient as well as the bundle methods. If checking the openness of the interiors of the candidate $\bar{P}_\sigma$ appears too laborious, one can employ the technique of *polynomial escape* in a given preferred direction $\mathbf{d}_1 \in \mathbb{R}^n$. After complementing it with $(n-1)$ directions $\mathbf{d}_i$ for $i = 2 \ldots n$ such that $\det(\mathbf{d}_1 \ldots \mathbf{d}_n) \neq 0$ one knows that for some σ and all small $0 < t \approx 0$

$$\mathbf{x}(t) \;=\; \mathring{\mathbf{x}} + \sum_{i=1}^{n} t^i \mathbf{d}_i \;\in\; P_\sigma$$

with P_σ being open. This corresponding σ and the corresponding gradient g_σ can be calculated independently of the parameter t using the function `firstsign` as described in [13]. It is a version of lexicographic differentiation introduced by Nesterov [34] and generalized to the case of composite functions by Khan and Barton [26]. The resulting g_σ is active in the direction $\mathbf{d}_1$ in

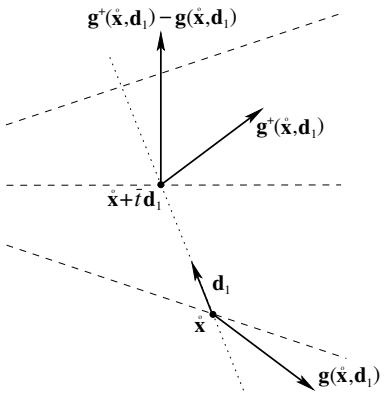

Fig. 4 Directional active gradients

that for some $\bar{t} > 0$

$$\mathbf{g}_\sigma^T \mathbf{d}_1 t \;=\; \Delta\varphi(\mathring{\mathbf{x}}; t\,\mathbf{d}_1) \;=\; \varphi(\mathring{\mathbf{x}} + t\mathbf{d}_1) - \varphi(\mathring{\mathbf{x}}) + \mathcal{O}(t^2) \quad \text{for} \quad 0 \le t \le \bar{t}\,.$$

The procedure for computing directionally active generalized gradients is actually matrix free and can be implemented efficiently using the so-called reverse mode of AD [16] except in the most degenerate cases. In those bad situations the complexity might equal that of the forward mode of AD, namely n times the complexity of evaluating the function $\varphi(\mathbf{x})$ itself, see also [25].

In the scenario $\varphi \in C^d_{\text{abs}}(\overline{\mathcal{D}})$ the directionally active gradient represents a little bit more then just any generalized gradient. The so-called *critical multiplier* gives important information about the nonsmoothness of φ from $\mathring{\mathbf{x}}$ in the direction $\mathbf{d}_1$. Moreover, if $\bar{t} < \infty$ we can also provide a gradient $\mathbf{g}^+(\mathring{\mathbf{x}}, \mathbf{d}_1)$ that is active in the direction $\mathbf{d}_1$ on the abs-linearization $\Delta\varphi(\mathring{\mathbf{x}}, \Delta\mathbf{x})$ at the point $\bar{t}\mathbf{d}_1$. The difference $\mathbf{g}^+(\mathring{\mathbf{x}}, \mathbf{d}_1) - \mathbf{g}(\mathring{\mathbf{x}}, \mathbf{d}_1)$ will then be a normal of the hyperplane separating the polyhedron before and beyond the kink location $\bar{t}\mathbf{d}_1$. This situation is depicted in Figure 4, where the dashed lines represent the kinks. Of course we can expect that on the underlying nonlinear function the situation is similar up to perturbations of $\mathcal{O}(\|\bar{t}\mathbf{d}_1\|^2)$. This should provide a lot of useful information for any kind of method based on generalized gradients.

Provided the complementing of $\mathbf{d}_1$ by $(n-1)$ linearly independent directions is continuous, the mapping

$$(\mathring{\mathbf{x}}, \mathbf{d}_1) \in \mathbb{R}^{n+n} \;\mapsto\; (\mathbf{g}, \bar{t}) \in \mathbb{R}^n \times (0, \infty)$$

has the following property: The multiplier $\bar{t}$ is continuous in the sense of extended real valued functions. The gradient $\mathbf{g}$ itself may have jumps and reduces to the Fréchet gradient, wherever that exists. The second gradient

$\mathbf{g}^+(\mathbf{x};\mathbf{d}_1)$ will be in many cases an ε-gradient at $\mathring{\mathbf{x}}$, but obviously one of them will generally not be enough to model φ locally even in the convex case. Of course, bundle methods could collect several of them, picking up several hyperplanes at a time.

$\varepsilon-$activity

Since the limiting and Clarke generalized differentials are not inner semicontinuous [35], the minimal norm of their elements gives no indication of the distance to any stationary point, in particular stopping criteria can not be based on it. Therefore, nonsmooth analysis has partly abandoned the strictly local point of view and introduced $\varepsilon-$differentials, which take note of the objective function behavior nearby. In our notation the definition of the Goldstein ε-differential reads

$$\partial^G_\varepsilon\varphi(\mathring{\mathbf{x}}) \;\equiv\; \mathrm{conv}\left\{\bigcup \partial^L\varphi(\mathbf{x}) : \mathbf{x}\in \bar{B}(\mathring{\mathbf{x}};\varepsilon)\right\}\;.$$

By definition the Goldstein ε-differential is inner semicontinuous. Obviously, we have $\partial\varphi(\mathring{\mathbf{x}}) = \partial^G_0\varphi(\mathring{\mathbf{x}})$. The natural question is how the Goldstein ε-differential $\partial^G_\varepsilon\varphi(\mathring{\mathbf{x}})$ can actually be computed. Of course, looking at all points $\mathbf{x}$ in a spherical neighborhood of the reference point $\mathring{\mathbf{x}}$ and computing the convex hull of the union of the limiting subdifferentials $\partial^L\varphi(\mathbf{x})$ appears practically impossible.

In the case of limiting gradients as in Equation (15) we only looked at signatures σ and their gradients $\mathbf{g}_\sigma$ of which we were certain that they are active for the abs-linearization and thus the function itself at $\mathring{\mathbf{x}}$. For all these neighboring signatures σ we know that $\sigma \succ \mathring{\sigma}$ which is equivalent to $\Sigma\mathring{\mathbf{z}} \geq 0$. That means when $\mathring{z}_i \neq 0$ the σ_i must have the same sign as $\mathring{\sigma}_i$ and where $\mathring{z}_i = 0$ we may choose freely $\sigma_i \in \{-1,1\}$. In order to get a larger set of gradients we may relax the condition on σ and only require that $\Sigma\mathring{\mathbf{z}} > -\varepsilon\,\mathbf{e}$ for the given $\varepsilon > 0$ and $\mathbf{e} = (1,1,\ldots,1)$. Then we define the corresponding limiting ε-differential

$$\partial^L_\varepsilon\varphi(\mathring{\mathbf{x}}) \;=\; \left\{\mathbf{a}+\mathbf{b}^T(\Sigma-L)^{-1}Z : \Sigma\mathring{\mathbf{z}} > -\varepsilon\,\mathbf{e}\right\} \;\supset \partial^L_0\varphi(\mathring{\mathbf{x}}) \;\supset\; \partial^L\varphi(\mathring{\mathbf{x}})\;. \quad (16)$$

and correspondingly "our" $\varepsilon-$differential simply as

$$\partial_\varepsilon\varphi(\mathring{\mathbf{x}}) = \mathrm{conv}(\partial^L_\varepsilon\varphi(\mathring{\mathbf{x}})).$$

Now we establish the desirable inner semicontinuity of both.

Lemma 2. *For fixed $\varepsilon > 0$ the multi-function $\mathbf{x} \rightrightarrows \partial^L_\varepsilon\varphi(\mathbf{x}) \subset \mathbb{R}^n$ and its convex hull $\mathbf{x} \rightrightarrows \partial_\varepsilon\varphi(\mathbf{x}) = \mathrm{conv}\{\partial^L_\varepsilon\varphi(\mathbf{x})\} \subset \mathbb{R}^n$ are inner semicontinuous.*

Proof. First let us consider any $\mathring{\mathbf{g}} \in \partial^L_\varepsilon\varphi(\mathring{\mathbf{x}})$ and its corresponding Σ satisfying

$$\mathring{\mathbf{g}} \;=\; \mathring{\mathbf{a}}^T + \mathring{\mathbf{b}}^T(\Sigma - \mathring{L})^{-1}\mathring{Z} \quad \text{and} \quad \Sigma\mathring{\mathbf{z}} > -\varepsilon\,\mathbf{e}$$

with σ definite and thus $|\det(\Sigma)| = 1$ without loss of generality. Moreover consider any sequence $\mathbf{x}_k \to \mathring{\mathbf{x}}$ and the corresponding $\mathbf{z}_k \to \mathring{\mathbf{z}}$. Then we must have by continuity also $\Sigma\mathbf{z}_k \to \Sigma\mathring{\mathbf{z}} > -\varepsilon\,\mathbf{e}$ so that already $\Sigma\mathbf{z}_k > -\varepsilon\,\mathbf{e}$ for all large k. Thus the corresponding $\mathbf{g}_k \;=\; \mathbf{a}_k + \mathbf{b}_k^T(\Sigma - L_k)^{-1}Z_k$ belong to $\partial_\varepsilon^L\varphi(\mathbf{x}_k)$ and of course their limit is $\mathring{\mathbf{g}}$. Thus every element of $\partial_\varepsilon^L\varphi(\mathring{\mathbf{x}})$ is the limit of limiting ε-gradients at any sequence converging to $\mathring{\mathbf{x}}$. Any $\mathring{\mathbf{g}} \in \partial_\varepsilon\varphi(\mathring{\mathbf{x}})$ is a convex combination of at most $n+1$ elements of $\partial_\varepsilon^L\varphi(\mathring{\mathbf{x}})$. As we have shown above each one of them is the limit of elements of $\partial_\varepsilon^L\varphi(\mathbf{x}_k)$ for any given sequence $\mathbf{x}_k \to \mathring{\mathbf{x}}$. Their convex combinations with the same coefficients belong to $\partial_\varepsilon\varphi(\mathbf{x}_k)$, which completes the proof. □

The lemma implies in particular that if any sequence $\mathbf{x}_k$ converges to a point $\mathbf{x}$ that is ε-stationary, the smallest elements $\operatorname{short}(\partial_\varepsilon\varphi(\mathbf{x}_k))$ of the $\partial_\varepsilon\varphi(\mathbf{x}_k)$ with respect to the Euclidean norm must converge to 0. Hence any stopping criterion $\|\operatorname{short}(\partial_\varepsilon^G\varphi(\mathbf{x}_k))\| < \delta$ for some positive δ must eventually be satisfied. Let us look at the situation in case of the half pipe function as defined in Equation (1) at origin $\mathring{\mathbf{x}} = 0$, where we have

$$\begin{aligned}\partial^K\varphi(0,0) = \{(0,0)\} &\subset \partial^L\varphi(0,0) = \{(-1,0),(0,0)\}\\ &\subset \liminf_{\varepsilon\to 0}\partial_\varepsilon^G\varphi(0,0) = \liminf_{\varepsilon\to 0}\operatorname{conv}\{(-1,0),(0,2\,x_2) : |x_2| < \varepsilon\}\ .\end{aligned}$$

Here, the Goldstein ε-subdifferential can be computed exactly but that is a very special situation. From now on we only consider *our* limiting ε-differential at a particular convergent sequence.

If we had a sequence $\mathbf{x}_k$ that converges to $0 \in \mathbb{R}^2$ all the time staying in the quadratic crescent $\mathcal{S}_{1,1}$ where $0 < x_{k,1}$ and $x_{k,2}^2 > x_{k,1}$ then we have for all k the singleton

$$\partial_\varepsilon^L\varphi(\mathbf{x}_k) = \{(-1, 2\,x_{k,2})\} = \partial_\varepsilon^G\varphi(\mathbf{x}_k) = \operatorname{short}(\partial_\varepsilon^G\varphi(\mathbf{x}_k))\ .$$

Then the length of $\operatorname{short}(\partial_\varepsilon^G\varphi(\mathbf{x}_k))$ would stay constantly greater than 1 despite the convergence to the Clarke stationary and even critical point $\mathring{\mathbf{x}} = 0$. On the other hand since $\mathbf{z}_k \to 0$, the condition $\Sigma\mathbf{z}_k > -\varepsilon\,\mathbf{e}$ is eventually satisfied for all $\sigma \in \{-1,1\}^2$ so that the late $\partial_\varepsilon^L\varphi(\mathbf{x}_k)$ are given by

$$\begin{aligned}\partial_\varepsilon^L\varphi(\mathbf{x}_k) \;&=\; \left\{\mathbf{a}_k + \mathbf{b}^T\begin{bmatrix}\sigma_1 & 0\\ \frac{1}{2} & \sigma_2\end{bmatrix}^{-1}\begin{bmatrix}1 & 0\\ -\frac{1}{2} & 2x_{k,2}\end{bmatrix}\middle|\ \sigma \in \{-1,1\}^2\right\}\\ &=\; \left\{\mathbf{a}_k + \mathbf{b}^T\begin{bmatrix}\sigma_1 & 0\\ -\frac{1}{2}\sigma_1\sigma_2 & \sigma_2\end{bmatrix}\begin{bmatrix}1 & 0\\ -\frac{1}{2} & 2x_{k,2}\end{bmatrix}\middle|\ \sigma \in \{-1,1\}^2\right\}\\ &=\; \left\{\mathbf{a}_k + \mathbf{b}^T\begin{bmatrix}\sigma_1 & 0\\ -\frac{1}{2}\sigma_2(\sigma_1+1) & 2\,x_{k,2}\sigma_2\end{bmatrix}\middle|\ \sigma \in \{-1,1\}^2\right\}\ .\end{aligned}$$

These are actually four different generalized gradients. However as we let $x_{k,2}$ tend to zero we get

$$\lim_{k\to\infty} \partial_\varepsilon^L \varphi(\mathbf{x}_k) =$$
$$(-0.25\ 0) + \left\{ (-0.25\ 0.5) \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}, (-0.25,\ 0.5) \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, (-0.25,\ 0.5) \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix} \right\}$$

so that we get only the two generalized gradients

$$(-0.25,\ 0) + (-0.75,\ 0) = (-1\ 0) \quad \text{and} \quad (-0.25,\ 0) + (0.25,\ 0) = (0,\ 0)\ .$$

Hence, in this case we have that the limiting differential is indeed contained and thus equal to the inner limit of our limiting ε-differential as a sequence of points converging to the limit $\mathring{\mathbf{x}}$. The same relation applies to their convex hulls, which we simply call ε-differentials.

Beyond mere gradient activity

The property $0 \in \mathrm{conv}\{\partial^K \varphi(\mathring{\mathbf{x}})\}$ of conic stationarity is considerably more restrictive than that of Clarke stationarity, i.e., $0 \in \mathrm{conv}\{\partial^L \varphi(\mathring{\mathbf{x}})\}$, which in turn is more restrictive than $0 \in \partial_\varepsilon^G \varphi(\mathring{\mathbf{x}})$. However, all are merely depending on which gradients are active at some arbitrary points near the reference point $\mathring{\mathbf{x}}$. The relative positions do not matter, which is why $|x|$ and $-|x|$ have the same limiting gradient $\{-1, +1\}$ and generalized gradient as well as ε-differential, namely $[-1, 1]$, respectively. In this case, there is no difference between the conic, Clarke and Goldstein ε-differential. Obviously that is not very useful in the context of optimization, where one wants to distinguish between minimizers and maximizers. To do that one must look at a proper local model function.

6 Checking Criticality and Second Order Optimality

It is immediately clear from Equation (10) that $\mathbf{x}_* = \mathring{\mathbf{x}}$ can only be a local minimizer of φ if it is a local minimizer of $\Delta\varphi(\mathring{\mathbf{x}}; \Delta\mathbf{x})$ with respect to $\Delta\mathbf{x} \approx 0$. We call that first order minimality (FOM). It is not difficult to see that on the function class $C^d_{\mathrm{abs}}(\overline{\mathcal{D}})$ this property is equivalent to criticality as defined in [1] and [2], where $0 \in \mathbb{R}^n$ must be a Fréchet subgradient. The term "criticality" insinuates that critical points of φ should also be critical points of $-\varphi$, which is decidedly not the case. By the sign change first order minimal points turn into first order maximal points, which unfortunately yields the same acronym FOM. So the proper terminology remains to be decided upon.

In general, i.e., for functions outside $C^d_{\rm abs}(\overline{\mathcal{D}})$ we know of no practical procedure to check a candidate point $\mathring{\mathbf{x}}$ for local minimality. Inside $C^d_{\rm abs}(\overline{\mathcal{D}})$ that can be done by a simple analysis of the abs-linear approximation data $Z, L, \mathbf{a}, \mathbf{b}, \mathbf{c}$ with $\mathbf{c} = \mathring{\mathbf{y}}$ obtained from an extended AD tool in the following way. For simplicity, we assume that for a given point $\mathbf{x}_*$ one has $z_i(\mathbf{x}_*) = 0$ for $i = 1 \ldots s$. This condition can be relaxed which requires technical reformulations [17]. For this reason we just concentrate on the simple case of full activity also called the localized case. Then one could first check, whether Z has full rank yielding LIKQ. As proven in the same paper, then first order optimality requires that for such a given point $\mathbf{x}_*$, there exists a Lagrange multiplier vector $\lambda_* \in \mathbb{R}^s$ such that

$$\mathbf{a}^T(\mathbf{x}_*, 0) + \lambda_*^T Z(\mathbf{x}_*, 0) = 0 \qquad \text{Tangential Stationarity (TS)} \tag{17}$$

$$F(\mathbf{x}_*, 0) = 0 \qquad \text{Full kink activity} \quad \text{and} \tag{18}$$

$$\mathbf{b}^T(\mathbf{x}_*, 0) + \lambda_*^T L(\mathbf{x}_*, 0) \geq |\lambda_*|^T \qquad \text{Normal Growth (NG)} \tag{19}$$

holds. Similar results apply if $\mathbf{x}_*$ is not localized in that some of the z_i are nonzero. It is important to note that these optimality conditions can be verified in polynomial time. If they do not hold, it is possible to construct a descent direction from the available derivative information $Z, L, \mathbf{a}, \mathbf{b}, \mathbf{c}$ as described in [17]. If all component inequalities hold strictly we say that Equation (19) represents strict normal growth.

First order minimality can be ensured for cluster points of the so-called proximal iteration

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \operatorname{argmin}_{\Delta \mathbf{x}} \left\{ \varphi(\mathbf{x}_k + \Delta \mathbf{x}) + \frac{\Delta q}{2} \|\Delta \mathbf{x}\|^2 \right\} . \tag{20}$$

Here, Δq can be any positive constant or vary within some interval. The practical challenge for the proximal point concept is that the inner problem of minimizing the right hand side seems almost as hard as the direct minimization of φ.

Before we develop an approximate version where $\varphi(\mathbf{x}_k + \Delta \mathbf{x})$ is replaced by the more tractable $\Delta\varphi(\mathbf{x}_k; \Delta \mathbf{x})$ let us briefly look at second order necessary and sufficiency conditions.

Second order piecewise differentiation and conditions

Here, we assume that $\varphi \in C^d_{\rm abs}(\overline{\mathcal{D}})$ with $d \geq 2$ so that all second order derivatives of F and f are continuous on the respective domains. These derivatives are conventional except that they are only valid on certain subspaces in certain polyhedral domains. We therefore talk of second order piecewise differentiation. Abs-linearization is a form of first order piecewise differentiation but it is more powerful in it works out the polyhedral decomposition at the same

time, which is then relevant for higher order piecewise differentiation. As of now we believe that differentiation on nonpolyhedral domains is impractical. The equalities in our first order condition represent $n + s$ equations in the unknowns $(\mathbf{x}_*, \lambda_*)$ whose Jacobian is given by the saddle point matrix

$$\begin{bmatrix} H & Z^T \\ Z & 0 \end{bmatrix} \in \mathbb{R}^{(n+s)\times(n+s)} \tag{21}$$

with

$$H = H(\mathbf{x}_*, \lambda_*) \equiv f(\mathbf{x}_*, 0)_{\mathbf{xx}} + \left(\lambda^T F(\mathbf{x}_*, 0)\right)_{\mathbf{xx}} \in \mathbb{R}^{n\times n} .$$

Obviously the Hessian H is the second derivative of the Lagrangian

$$\mathcal{L}(\mathbf{x}, 0, \lambda) = f(\mathbf{x}, 0) + \lambda^T F(\mathbf{x}, 0) \tag{22}$$

with respect to $\mathbf{x}$. The saddle point Jacobian (21) is nonsingular provided we have second order sufficiency in that $U^T H U \succ 0$, where the columns of $U \in \mathbb{R}^{n\times(n-s)}$ span the null space of Z. Then we have a sufficient optimality condition in combination with tangential stationarity and strict normal growth. If $\det(U^T H U) = 0$, but the projected Hessian is still positive semidefinite we have a second order necessary condition of optimality. Wherever LIKQ holds the function φ will be smooth within P_σ but may have kinks (upward or downward) along certain normal directions. As was proven in [17] this geometry corresponds to the $\mathcal{VU}$ decomposition of Mifflin and Sagastizábal [33] and Lewis [29], where the kinks are restricted to point upward. Here we can define at a point $\mathbf{x}$ the pair of orthogonal subspaces

$$\mathcal{U}(\mathbf{x}) \equiv \text{range}\,(U(\mathbf{x})) \quad \text{and} \quad \mathcal{V}(\mathbf{x}) \equiv \mathcal{U}(\mathbf{x})^T.$$

It should be noted that the $\mathcal{VU}$ decomposition exists for some functions outside $C^d_{\text{abs}}(\overline{\mathcal{D}})$, for example again the Euclidean norm in two variables or more.

Reaching Criticality

Based on the abs-linearisation described at the end of Section 3, the following iterative optimization algorithm was proposed in [13]

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \text{argmin}_{\,\Delta\mathbf{x}} \left\{\varphi(\mathbf{x}_k) + \Delta\varphi(\mathbf{x}_k; \Delta\mathbf{x}) + \tfrac{q}{2}\|\Delta\mathbf{x}\|^2\right\} . \tag{23}$$

We call this approach SALMIN for Successive Abs-Linear MINimization. The penalty factor q of the quadratic term is an estimated bound on the discrepancy between φ and its local abs-linear model given by

$$\varphi(\mathbf{x}_k) + \Delta\varphi(\mathbf{x}_k; \Delta\mathbf{x}) .$$

This method was shown in [13] to generate a sequence of iterates $(\mathbf{x}_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ whose cluster points are first order minimal. If the inner problem of minimizing the regularized piecewise linear model is not solved exactly, but increments $\Delta\mathbf{x}$ that are merely Clarke stationary for $\Delta\varphi$ are accepted also, then the cluster points are guaranteed to be also Clarke stationary as shown in [9].

The SALMIN algorithm as stated in Equation (23) can be interpreted also as a quadratic overestimation method, where the error between the model and the real function is bounded by a power of the distance, see, e.g., [11, 14]. This approach is in some sense related to a proximal point method as stated in Equation (20). However, in Equation (23) the local abs-linear model of the function to be minimized at the current iterate $\mathbf{x}_k$ is used instead of the original function. This makes the solution of the inner optimization problem considerably easier in comparison to the proximal point method. Moreover, without looking at generalized gradients or $\varepsilon-$subdifferentials SALMIN has a very simple stopping criterion. The outer iteration terminates as soon as the objective function reduction promised by the solution of the inner problem falls below a user supplied tolerance.

One possible strategy to solve the inner problem, i.e., determine the minimizer of

$$\operatorname{argmin}_{\Delta\mathbf{x}} \left\{ \varphi(\mathbf{x}_k) + \Delta\varphi(\mathbf{x}_k; \Delta\mathbf{x}) + \tfrac{q}{2}\|\Delta\mathbf{x}\|^2 \right\} ,$$

exploits the polyhedral domain decomposition defined by Equation (12). Starting with an arbitrary initial point and the corresponding polyhedron, one can derive an adapted QOP solver by exploiting the local first order optimality condition from [17] as stated for the localized case (18) in Equations (17) and (19). This strategy is based on the computation of stationary points by successively activating and dropping kinks appropriately as described in detail in [18].

7 Demonstration on Crescent

The various quantities that we promised as benefits of piecewise differentiation and the convergence behavior of SALMIN are illustrated on the two dimensional Crescent example [3, Nr. 21 in Sec. 9.1], namely

$$y = f(x_1, x_2) \;=\; \max\{x_1^2 + (x_2 - 1)^2 + x_2 - 1, -x_1^2 - (x_2 - 1)^2 + x_2 + 1\}$$

with the starting point $(-1.5, 2)$. In abs-normal form we can write

$$z_1 = F(x_1, x_2) = x_1^2 + (x_2 - 1)^2 - 1$$

and

$$y = f(x_1, x_2, |z_1|) = x_2 + |z_1| \; .$$

The new form was achieved by replacing $\max(u, w)$ with the equivalent value $\frac{1}{2}[u+w+\mathrm{abs}(w-u)]$ and then canceling various terms, which can of course be done by computer algebra or an AD package. Here one sees immediately that the set of kink locations is formed by the shifted unit circle $x_1^2+(x_2-1)^2 = 1$.

Abs-linear Approximation

With respect to the abs-linear form we note that since there is only one switching variable we must have the trivial strictly lower triangular matrix $L = 0 \in \mathbb{R}^{1\times 1}$. The remaining parts of the abs-linear form at a point $\mathring{\mathbf{x}}$ are given by

$$\begin{aligned} &Z = \partial_{\mathbf{x}} z_1 = 2(\mathring{x}_1,\ \mathring{x}_2 - 1),\ a = (0,\ 1)^T,\ b = 1 \ \text{ and} \\ &c = \mathring{z} = \mathring{x}_1^2 + (\mathring{x}_2 - 1)^2 - 1\ . \end{aligned}$$

The matrix Z has full rank except at the center $(0, 1)$ of the circle. Hence, LIKQ is satisfied everywhere on the kink circle.

Looking for optimal points

To test for optimality we first look at tangential stationarity, which requires that

$$0 = (0, 1) + \lambda\, 2(x_1, x_2 - 1) \quad \text{and} \quad z_1 = 0\,.$$

This system of equations has the two solutions solution $(x_1, x_2) = (0, 0)$ with $\lambda = \frac{1}{2}$ and $(x_1, x_2) = (0, 2)$ with $\lambda = -\frac{1}{2}$. The normal growth requires that $b = \partial y/\partial|z_1| \equiv 1 \geq |\lambda| = \frac{1}{2}$ which is satisfied as strict inequality at both points. Thus we have at both points first order optimality, which is also known as criticality. Finally, at $\mathbf{x} = (0, 2)$ the null space of $Z = (0, 2)$ is spanned by $U = (1, 0)^T$ so that the Hessian of the Lagrangian at the first order optimal point $\mathbf{x} = (0, 2)$ with $\lambda = -\frac{1}{2}$

$$-\tfrac{1}{2}(1,\ 0) \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \ = \ -1 \ < \ 0\ .$$

Here we have used that f is linear and hence its second derivatives vanish completely. Thus, the first order optimal point $(0, 2)$ does not satisfy the second order necessary condition and cannot be a minimizer. At the only other point satisfying tangential stationarity, namely the origin, we have $Z = (0, –2)$ so that with $U = (1, 0)^T$ and the positive Lagrange multiplier one obtains

$$\tfrac{1}{2}(1,\ 0) \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \ = \ 1 \ > \ 0\ .$$

This projected Hessian is positive definite and the origin is therefore a strict local minimizer and thus in fact the one and only global minimizer. Notice that the $\mathcal{VU}$ decomposition is well defined all around the kink circle with $\mathcal{V}$ being the radial direction, i.e., the normal of $z_1 = 0$ and $\mathcal{U}$ the tangential direction. Everywhere the kink is pointed upwards, although that need nor be valid in general.

Let us go back and calculate the other goodies at some general point $\mathring{\mathbf{x}}$, say the usual starting point $\mathring{\mathbf{x}} = (\mathring{\mathbf{x}}_1, \mathring{\mathbf{x}}_2) = (-1.5, 2)$. There we have $\mathring{z} = \frac{9}{4}$ and thus $\mathring{\sigma} = 1 = \Sigma$. Moreover, $\mathring{Z} = (-3,\ 2)$ so that independently of any preferred direction differentiation yields the gradient

$$\mathring{\mathbf{g}} = \mathbf{a}^T + \mathbf{b}(\Sigma - L)^{-1} Z = (0,1) + 1 \cdot (-3, 2) = (-3, 3) \quad \Longrightarrow \quad -\mathring{g} = (3, -3)\ .$$

With respect to the limiting ε-subdifferential, according to Equation (16) for $\varepsilon > 0$ we are admitting $\sigma \in \{-1, 1\}$ that satisfy $\sigma \frac{9}{4} > -\varepsilon$. Hence, $\sigma = -1$ and the corresponding

$$g_\sigma = (0,\ 1) + (1)(-1)(-3,\ 2) = (3,\ -1)$$

will only be an ε-gradient when $\varepsilon > \frac{9}{4}$. Obviously that is a strong condition but then the reference point $(-1.5, 2)$ is quite some distance away from the next kink.

Finally, let us consider the performance of our SALMIN approach from the standard starting point with $q = 3$ constant. As one can see on the right hand side of Figure 5 the convergence rate is clearly linear. It can not be better because no effort is made by SALMIN to approximate the curvature term that defines the circular valley.

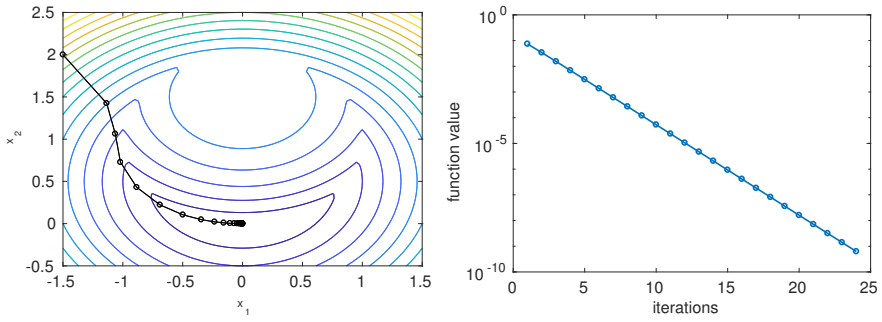

Fig. 5 Iterates and function values of SALMIN on Crescent example

We have applied earlier versions of SALMIN to most of the academic problems listed in [3]. The results in [9] are quite competitive with a generalization of BFGS [30] and the bundle method [24]. In fact the number of outer iterations is usually significantly smaller and a thorough comparison

of the runtime cost of solving the inner problem remains to be done. In any case the inner loop of SALMIN is still undergoing rapid development, especially in view of larger dimensional applications. Another generalization that is under way is the extension to problems with constraints, which may be of complementarity type.

8 Covering the Euclidean norm

In the context of geometric modeling, see, e.g., [32], one may easily think of optimization problems or systems of constraints that involve for $\mathbf{u} \in \mathbb{R}^k$ the Euclidean norm

$$||\mathbf{u}|| \;=\; \left(\sum_{i=1}^{k} u_i^2\right)^{\frac{1}{2}} \;=\; \|(u_1, \|(u_2, u_3, \ldots, u_k)\|)\| \;. \tag{24}$$

The identity on the right shows that the Euclidean norm in $k > 2$ variables can be expressed recursively in terms of the binary Euclidean norm $\|(u_1, u_2)\| = \sqrt{u_1^2 + u_2^2}$. This elemental function is of course a generalization of our beloved unary absolute value $\|u\| = |u|$ for $u \in \mathbb{R}$. Already the binary Euclidean norm is no longer piecewise differentiable, because at the origin one would need more than finitely many C^1 selection functions to represent it. However it is Lipschitz continuous with constant 1 and almost everywhere differentiable, which one can see directly without referring to Rademacher. As we already foreshadowed at the end of Section 3 we now consider the extension of $C^d_{\rm abs}(\overline{\mathcal{D}})$ obtained by allowing not only the univariate abs$(\cdot)$ but its multivariate generalization $\|\cdot\|$.

Problems in $C^d_{\rm euc}(\overline{\mathcal{D}})$

As main example we consider the simplest so-called location problem [19], which goes back to Fermat in the planar case. Given m distinct *client* points $\mathbf{y}_j \in \mathbb{R}^k$ for $j = 1, \ldots, m$ we are looking for a *supply* point $\mathbf{x}$ that minimizes the sum of the Euclidean distances to the clients.

$$\min \varphi_m(\mathbf{x}) \;\equiv\; \sum_{j=1}^{m} \|\mathbf{x} - \mathbf{y}_j\| \;\in C^\infty_{\rm euc}(\mathbb{R}^n, \mathbb{R}) \;.$$

The problem is convex, coercive and Lipschitz continuous so that it must have a nonempty compact and convex solution set. Moreover $\varphi_m(\mathbf{x})$ is also differentiable except where $\mathbf{x} = \mathbf{y}_j$ for some $1 \le j \le m$.

When $m = 3$ and thus w.l.o.g $n = 2$ we have the classical Fermat problem, whose solution was be constructed geometrically with a pair of compasses and ruler by Toricelli. Now suppose one has solved the problem in the horizontal plane, i.e., for three points and their geometric median

$$\mathbf{y}_1 = (y_{1,1}, y_{1,2}, 0),\ \mathbf{y}_2 = (y_{2,1}, y_{2,2}, 0),\ \mathbf{y}_3 = (y_{3,1}, y_{3,2}, 0),\ \mathbf{w} = (w_1, w_2, 0)\ .$$

The minimizer $\mathbf{w}$ is the only stationary point of $\varphi_3(.)$ so that one has

$$\nabla\varphi_3(\mathbf{w}) = 0 \quad \text{and} \quad \varphi_3(\mathbf{x}) = \varphi_3(\mathbf{w}) + \mathcal{O}(\|\mathbf{x} - \mathbf{w}\|^2)\ .$$

Now let us add a forth data point $\mathbf{y}_4 = (y_{4,1}, y_{4,2}, y_{4,3})$ that is reasonable close to $\mathbf{w}$. Then we will have for the new $\varphi_4(.)$ that

$$\varphi_4(\mathbf{x}) \ = \ \varphi_3(\mathbf{x}) + \|\mathbf{x} - \mathbf{y}_4\| \ = \ \varphi_3(\mathbf{w}) + \|\mathbf{x} - \mathbf{y}_4\| + \mathcal{O}(\|\mathbf{x} - \mathbf{w}\|^2)$$

and in particular when $\mathbf{y}_4 = \mathbf{w}$

$$\varphi_4(\mathbf{x}) \ = \ \varphi_3(\mathbf{w}) + \|\mathbf{x} - \mathbf{w}\| + \mathcal{O}(\|\mathbf{x} - \mathbf{w}\|^2)$$

with $\mathbf{x} = \mathbf{w}$ as the nonsmooth (global) minimizer of $\varphi_4(.)$. Moreover, since for $\mathbf{x} \neq \mathbf{y}_4$

$$\nabla\varphi_4(\mathbf{x}) \ = \ \nabla\varphi_3(\mathbf{x}) + (\mathbf{x} - \mathbf{y}_4)/\|\mathbf{x} - \mathbf{y}_4\|$$

the same will be true for all $\mathbf{y}_4$ with $\|\nabla\varphi_3(\mathbf{y}_4)\| < 1$ and even $\|\nabla\varphi_3(\mathbf{y}_4)\| = 1$ since $\varphi_4(.)$ is convex. For each of these $\mathbf{y}_4$ we have a convex test problem with the global minimizer $\mathbf{x} = \mathbf{y}_4$, at which $\varphi_4(\mathbf{x})$ is dominated by the Euclidean norm and thus not differentiable.

A very similar situation arises in *compressive sensing* [10] where the distance $\|\mathbf{x}-\mathbf{y}_4\|$ of the variable vector $\mathbf{x}$ to a base point (often $\mathbf{y}_4 = 0$) is added to a smooth residual, here $\varphi_3(\mathbf{x})$. Then the base point is the *sparse* global minimizer as long as the smooth part is comparatively stationary. Only when the smooth part is as steep as the flanks of the norm term the minimizer can be pulled away from the reference points. Now the question arises how this type of problem can be minimized algorithmically. The simplest possible test problem in two variables would be

$$\varphi(x_1, x_2) \ = \ \sqrt{x_1^2 + x_2^2} + (\lambda, 0)^T x \quad \text{with} \qquad |\lambda| \leq 1 \quad .$$

We have expressed the Euclidean norm implicitly and assume at first that it will not be recognizable to the optimization algorithm. Lewis and Overton [30] have called this problem the *tilted norm* function, which happens to be the only situation for which they can prove and not only observe the convergence of their BFGS method with a special line search. From a starting point with $x_2 = 0$, steepest descent with any kind of line search will behave exactly as on the univariate problem $|x| + \lambda x$. In our experience steepest descent with a Armijo type line-search stalls completely in the vicinity of

the optimizer. Lewis and Overton have shown theoretically that in combination with their specially for nonsmooth problems adapted Armijo line-search, steepest descent exhibits a sublinear convergence rate in terms of the number of function evaluations.

In the one dimensional case with $\mathrm{abs}(x) = |\cdot|$ identified as such, our SALMIN approach would of course yield convergence from any initial point in one step. On the two dimensional problem without any hint of nonsmooth elementals it would behave like steepest descent with the coefficient q being incremented several times in each line search. Theoretically the fact of convergence can be deduced by contradiction as follows. If there was a ball about $\mathbf{y}_4$ which was not reached by anyone of the iterates one could modify the convex function $\varphi_4(.)$ inside such that the cone singularity is smoothened out. Then our standard convergence theory would ensure convergence into the ball yielding a contradiction. Note that the other three points of nondifferentiability have a much higher function value and can therefore not be approached if the iteration is started below. Also, notice that we have assumed throughout like in [30] that the single point of nondifferentiability, i.e., the global minimizer itself is never reached exactly by any iterate. Of course a small fixed stepsize as is popular in machine learning will ultimately lead to oscillations back and forth across the base point. One might argue that the solution error may then be quite small during this chattering, but the whole purpose of these terms is to drive them exactly to zero and thus to achieve data sparsity.

So, en passant, we reach the tentative conclusion that on machine learning problems similar to Lasso [37] steepest descent converges sublinearly with line-search and does not converge at all for a fixed stepsize. Obviously, some thing needs to be done to overcome this impasse.

Clipped Root Linearization

We have seen in the previous section that approximating the Euclidean norm by its tangent plane (and equivalently the square root by its tangent line) does not yield good results on the kind of optimization problems in $C^d_{\mathrm{euc}}(\overline{\mathcal{D}})$. As it turns out the two approximation tasks are intimately related and by simply making a small modification to the root linearization we obtain a desired effect for the Euclidean root. Therefor we will go backward and start with the root, whose normal incremental linearization is given by

$$v = \sqrt{u} \quad \Longrightarrow \quad v + \Delta v = v + 0.5\Delta u/v \quad \Longleftrightarrow \quad \Delta v = 0.5\,\Delta u/v \tag{25}$$

with the tacit assumption that u and thus v are not exactly equal to zero. This propagation happens automatically under the rug when piecewise linearization is applied to a function evaluation procedure $y = \varphi(\mathbf{x})$. The value u and the increment Δu of the right hand side are computed from $\mathbf{x}$ and $\Delta\mathbf{x}$

via the preceding intermediate operations and Δv is the resulting increment of the left hand side. In other words the root is treated like any other differentiable univariate function, namely replaced by its tangent line. In contrast to the root itself, which is undefined for negative values, the linearization reaches arbitrarily large negative values. Therefore, one might argue that the user should be alerted in some way to the qualitative change for negative increments Δu much bigger than $u > 0$. The simple idea of taking the absolute value of the linear prediction leads to:

Definition 5 (Clipped Linearization). The *Clipped Linearization* of the root is given by

$$\begin{aligned} v = \mathrm{abs}(\sqrt{u}) \quad &\Longrightarrow \quad v + \Delta v = |v + 0.5\Delta u/v| \\ &\Longleftrightarrow \quad \Delta v = |v + 0.5\Delta u/v| - v \ . \end{aligned} \tag{26}$$

Furthermore, we will call this technique of maintaining non-negativity or other bounds of the original elemental by its piecewise linearization as *clipping*.

Of course, while maintaining characteristic properties of the original elemental we have introduced an extra kink and thus made the piecewise linear model a bit more complicated. However, we will assume that there are lots of kinks anyhow so that a few more do not make a significant difference. The linear and clipped approximation of the square root are depicted in Figure 6.

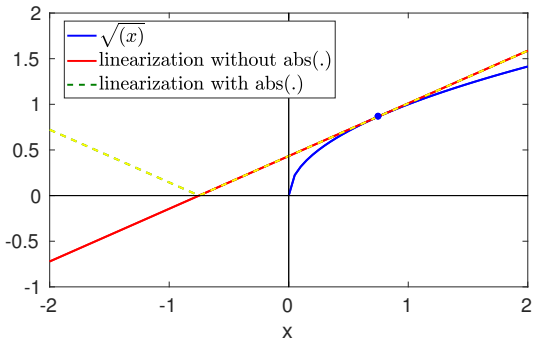

Fig. 6 Two different linearizations for the square root at $\mathring{\mathbf{x}} = 0.75$

The straight tangent line has been replace by a V-shaped line touching the horizontal axis at $\Delta u = -2v$. The nice thing here is that one does not have to change anything in the evaluation procedure except extending all $\sqrt{u}$ to $\mathrm{abs}(\sqrt{u})$, which is of course equivalent as far as the values themselves (but not the increments) are concerned. The piecewise linearization process by ADOL-C or some other abs-extended AD tool can then proceed as usual. Now the

question is what that mechanism does to the Euclidean norm. The usual differentiation of the Euclidean norm of $\mathbf{u} \in \mathbb{R}^k$ in the composite form (24) gives the linear approximation

$$v = ||\mathbf{u}|| \Longrightarrow v + \Delta v = (\mathbf{u} + \Delta \mathbf{u})^T \mathbf{u}/\|\mathbf{u}\| \quad \Longleftrightarrow \quad \Delta v = \Delta \mathbf{u}^T \mathbf{u}/\|\mathbf{u}\|$$

again tacitly assuming that $\mathbf{u} \neq 0$ and equivalently $v \neq 0$. Now if again we extend $\|\mathbf{u}\|$ to $\mathrm{abs}(\|\mathbf{u}\|)$ we get after some manipulations

$$\begin{aligned} v = \mathrm{abs}(\|\mathbf{u}\|) \quad &\Longrightarrow \quad v + \Delta v = |(\mathbf{u} + \Delta \mathbf{u})^T \mathbf{u}|/\|\mathbf{u}\| \\ &\Longleftrightarrow \quad \Delta v = |(\mathbf{u} + \Delta \mathbf{u})^T \mathbf{u}|/\|\mathbf{u}\| - v \; . \end{aligned} \tag{27}$$

This approximation of the Euclidean norm is a V-shaped valley whose bottom line is orthogonal to the reference point $\mathbf{u}$ as illustrated in Figure 7 Again

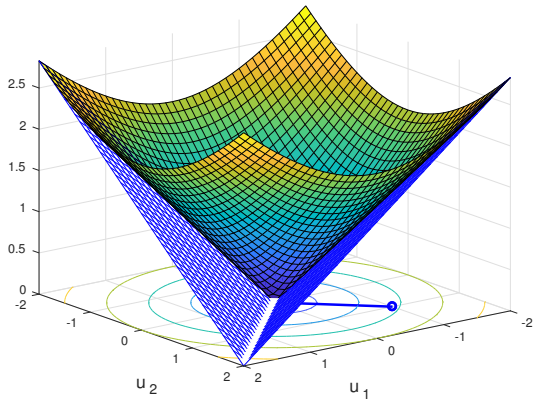

Fig. 7 The Euclidean norm $\|\mathbf{u}\|$ and its clipped linearization at $\mathring{\mathbf{u}} = (-1, 1)$

there is no need for any substantial recoding but simply one has to extend all $v = ||\mathbf{u}||$ to $v = \mathrm{abs}(\|\mathbf{u}\|)$ or even simpler use the expression (24) and extend $\sqrt{u}$ to $|\sqrt{u}|$ as suggested above.

We glossed a little bit about possible overflows when the scalar u or the vector $\mathbf{u}$ are small. Moreover, there is one important aspect that we have not mentioned. Namely the valley linearization of the Euclidean norm is not of second order and hence the generalized Taylor property (10) does no longer hold for the over-all abs-linearization. Of course, one might hope that that does not stop whatever algorithm one is using from converging, albeit at a possibly reduced rate. Specifically applying the current version of SALMIN without any modifications to the location problem also called Weber problem [40] as described above we get the linear convergence behavior displayed in Figure (8). The minimizer of $\varphi_3(.)$ with $\mathbf{y}_1 = (1, 1)$, $\mathbf{y}_2 = (-1, -1)$ and $\mathbf{y}_3 = (-1,)$ is $\mathbf{w} = (-0.577, 0.577, 0)$. Notice that the iteration appears to

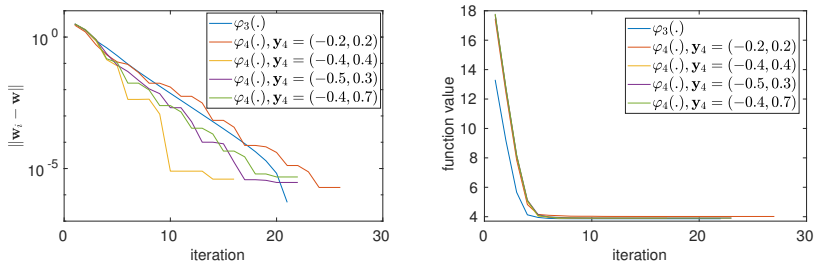

Fig. 8 Results of SALMIN with clipped square root for the Weber Problem

alternate between a step that barely reduces the distance to the solution, presumably moving along the bottom of the valley approximation and one that reduced the distance by a about a quarter. These numerical results appear quite satisfactory in view of the observation that normal descent methods are almost certain to have a sublinear rate, which in the presence of rounding errors means stalling not all that close to a solution. To the best of our knowledge the clipped versions (26) and (27) of the piecewise linearization of root and Euclidean norm have not yet appeared in the literature.

9 Summary and Conclusion

As indicated by the title we tried to sow some doubts regarding the plausibility of the popular "oracle" scenario, i.e., the availability of the function value and one generalized gradient. The key claim is that, if there is a way to compute a vector that is guaranteed to be a generalized gradient then one can apply piecewise differentiation and obtains lots of other goodies, like directionally active gradients, critical multipliers, approximating separating planes, conically active generalized gradients, active $\varepsilon-$gradients and the whole local abs-linear approximation in the form of two matrices and three vectors. That full local model naturally leads to the SALMIN method for which there is now an extensive theory [9] and [18]. The $\varepsilon-$active gradients defined by (16) were firstly introduced in this paper and their relation to the classical $\varepsilon-$differential of Goldstein deserves further exploration. They certainly have the advantage of practical computability with polynomial effort. Also the clipped linearisation as defined in Definition 5 for the root is proposed here for the first time.

It remains to be seen, which class of problems are efficiently treatable by piecewise differentiation or not. In the penultimate section we looked at the extension of $C^d_{\rm abs}(\overline{\mathcal{D}})$ to $C^d_{\rm euc}(\overline{\mathcal{D}})$ by generalization of the absolute value to the Euclidean norm in two and thus arbitrary many variables. It is found that the

piecewise linearization of the norm by a V-shaped valley rather than just its tangent plane appears very useful for avoiding the sublinear convergence of classical descent methods. The concept of abs-linearization is also extendable to reflexive Banach spaces and thus the optimization under PDE constraints. Finally let us remark that the abs-linear approximation can also be exploited for other fundamental numerical tasks like the solution of nonlinear systems and the integration of Lipschitz continuous dynamical systems.

References

1. Absil, P.A., Mahony, R., Andrews, B.: Convergence of the iterates of descent methods for analytic cost functions. SIAM Journal on Optimization **16**(2), 531–547 (2005)
2. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Mathematical Programming. Series B **116**(1-2), 5–16 (2009)
3. Bagirov, A., Karmitsa, N., Mäkelä, M.: Introduction to nonsmooth optimization. Theory, practice and software. Springer (2014)
4. Barton, P., Khan, K., Stechlinski, P., Watson, H.: Computationally relevant generalized derivatives: Theory, evaluation and applications. Optimization Methods and Software **33**(4–6), 1030–1072 (2018)
5. Bell, B.: CppAD. https://www.coin-or.org/CppAD/
6. Burke, J., Henrion, D., Lewis, A., Overton, M.: Stabilization via nonsmooth, nonconvex optimization. IEEE Transactions on Automatic Control **51**(11), 1760–1769 (2006)
7. Clarke, F.: Optimization and Nonsmooth Analysis. SIAM (1990)
8. Facchinei, F., Pang, J.S.: Finite-dimensional variational inequalities and complementarity problems. Vol. II. Springer (2003)
9. Fiege, S., Walther, A., Griewank, A.: An algorithm for nonsmooth optimization by successive piecewise linearization. Mathematical Programming, Series A (2018). DOI: 10.1007/s10107-018-1273-5
10. Foucart, S., Rauhut, H.: A mathematical introduction to compressive sensing. Birkhäuser/Springer (2013)
11. Griewank, A.: The modification of Newton's method for unconstrained optimization by bounding cubic terms. Tech. Rep. NA/12, University of Cambridge (1981)
12. Griewank, A.: Automatic directional differentiation of nonsmooth composite functions. In: Recent developments in optimization. 7th French-German conference on optimization, Dijon, France, June 27-July 2, 1994, pp. 155–169. Springer (1995)
13. Griewank, A.: On stable piecewise linearization and generalized algorithmic differentiation. Optimization Methods and Software **28**(6), 1139–1178 (2013)
14. Griewank, A., Fischer, J., Bosse, T.: Cubic overestimation and secant updating for unconstrained optimization of $C^{2,1}$ functions. Optimization Methods and Software **29**(5), 1075–1089 (2014)
15. Griewank, A., Streubel, T., Lehmann, L., Radons, M., Hasenfelder, R.: Piecewise linear secant approximation via algorithmic piecewise differentiation. Optimization Methods & Software **33**(4–6), 1108–1126 (2018)
16. Griewank, A., Walther, A.: Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation. SIAM (2008)

17. Griewank, A., Walther, A.: First and second order optimality conditions for piecewise smooth objective functions. Optimization Methods and Software **31**(5), 904–930 (2016)
18. Griewank, A., Walther, A.: Finite convergence of an active signature method to local minima of piecewise linear functions. Tech. rep., Universität Paderborn (2017). Available at optimization-online
19. Hamacher, H., Nickel, S.: Classification of location models. Location Science **6**(1–4), 229–242 (1998)
20. Hascoët, L., Pascual, V.: The Tapenade automatic differentiation tool: Principles, model, and specification. ACM Transactions on Mathematical Software **39**(3), 20:1–20:43 (2013)
21. Hegerhorst-Schultchen, L., Kirches, C., Steinbach, M.: On the relation between mpecs and optimization problems in abs-normal form. Tech. rep., Universität Hannover (2018). Available at optimization-online
22. Hintermüller, M.: Semismooth newton methods and applications. Tech. rep., Department of Mathematics, Humboldt-University Berlin (2010)
23. Hintermüller, M., Stadler, G.: A semi-smooth Newton method for constrained linear-quadratic control problems. ZAMM. Zeitschrift für Angewandte Mathematik und Mechanik **83**(4), 219–237 (2003)
24. Karmitsa, N., Mäkelä, M.: Limited memory bundle method for large bound constrained nonsmooth optimization: convergence analysis. Optim. Methods Softw. **25**(6), 895–916 (2010)
25. Khan, K.: Branch-locking ad techniques for nonsmooth composite functions and nonsmooth implicit functions. Optimization Methods and Software (2017). Published online: https://doi.org/10.1080/10556788.2017.1341506
26. Khan, K., Barton, P.: Evaluating an element of the Clarke generalized Jacobian of a composite piecewise differentiable function. ACM Transactions on Mathematical Software **39**(4), 28 (2013)
27. Khan, K., Barton, P.: A vector forward mode of automatic differentiation for generalized derivative evaluation. Optimization Methods and Software **30**(6), 1185–1212 (2015)
28. Klatte, D., Kummer, B.: Nonsmooth equations in optimization. Regularity, calculus, methods and applications. Kluwer Academic Publishers (2002)
29. Lewis, A.: Active sets, nonsmoothness, and sensitivity. SIAM Journal on Optimization **13**(3), 702–725 (2002)
30. Lewis, A., Overton, M.: Nonsmooth optimization via quasi-Newton methods. Mathematical Programming Series A **141**(1-2), 135–163 (2013)
31. Leyffer, S.: Mathematical programs with complementarity constraints. SIAG/OPTViews-and-News **14**(1), 15–18 (2003)
32. Liberti, L., Lavor, C., Maculan, N., Mucherino, A.: Euclidean distance geometry and applications. SIAM Review **56**(1), 3–69 (2014)
33. Mifflin, R., Sagastizábal, C.: A science fiction story in nonsmooth optimization originating at IIASA. Documenta Mathematica **Extra Vol.**, 291–300 (2012)
34. Nesterov, Y.: Lexicographic differentiation of nonsmooth functions. Mathematical Programming Series A **104**(2-3), 669–700 (2005)
35. Rockafellar, R., Wets, R.B.: Variational analysis. Springer (1998)
36. Scholtes, S.: Introduction to Piecewise Differentiable Functions. Springer (2012)
37. Tibshirani, R.: Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society. Series B **58**(1), 267–288 (1996)
38. Walther, A., Griewank, A.: Combinatorial Scientific Computing, chap. Getting Started with ADOL-C, pp. 181–202. Chapman-Hall CRC Computational Science (2012)
39. Walther, A., Griewank, A.: Characterizing and testing subdifferential regularity for piecewise smooth objective functions. Tech. Rep. SPP1962-038, Universität Paderborn (2017). Available at optimization-online
40. Weber, A.: Über den Standort der Industrien. J.C.B. Mohr (1909)