

First and second order optimality conditions for piecewise smooth objective functions

Andreas Griewank¹ and Andrea Walther²

March 7, 2016

Abstract

Any piecewise smooth function that is specified by an evaluation procedure involving smooth elemental functions and piecewise linear functions like min and max can be represented in the so-called abs-normal form. By an extension of algorithmic, or automatic, differentiation, one can then compute certain first and second order derivative vectors and matrices that represent a local piecewise linearization and provide additional curvature information. On the basis of these quantities we characterize local optimality by first and second order necessary and sufficient conditions, which generalize the corresponding KKT theory for smooth problems. The key assumption is the Linear Independence Kink Qualification (LIKQ), a generalization of LICQ familiar from nonlinear optimization. It implies that the objective has locally a so-called \mathcal{VU} decomposition and renders everything tractable in terms of matrix factorizations and other simple linear algebra operations. By yielding descent directions whenever they are violated the new optimality conditions point the way to a superlinearly convergent generalized QP solver, which is currently under development. We exemplify the theory on two nonsmooth examples of Nesterov.

Keywords: Abs-normal form, piecewise linearization, Karush-Kuhn-Tucker, second order optimality, projected Hessian, tangential stationarity, normal growth, \mathcal{VU} decomposition

¹School of Mathematical Sciences, Yachaytech, Urcuquí, Imbabura, Ecuador

²Institut für Mathematik, Universität Paderborn, Paderborn, Germany

1 Introduction and assumptions

According to [7] Nesterov suggested the three Rosenbrock like test functions $\varphi_\nu : \mathbb{R}^n \mapsto \mathbb{R}$ given by

$$\varphi_0(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} (x_{i+1} - 2x_i^2 + 1)^2, \quad (0)$$

$$\varphi_1(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |x_{i+1} - 2x_i^2 + 1|, \quad (1)$$

$$\varphi_2(x) = \frac{1}{4} |x_1 - 1| + \sum_{i=1}^{n-1} |x_{i+1} - 2|x_i| + 1|. \quad (2)$$

Only $\varphi_0(x)$ is smooth and the subscripts $\nu = 0, 1, 2$ for $\varphi_\nu(x)$ reflect what is called the switching depth of the nonsmoothness represented in abs-normal form [3]. Correspondingly, we will refer to the zeroth, first, and second example with the smooth, zeroth example playing a minor role. The two nonsmooth variants $\varphi_1(x)$ and $\varphi_2(x)$ were carefully investigated in [7]. All three functions have the unique global minimizer $x = (1, 1, \dots, 1) \in \mathbb{R}^n$ but the last one $\varphi_2(x)$ has $2^{n-1} - 1$ other stationary points, at which nonsmooth optimization algorithms may get stuck even though none of them is a local minimizer.

Here stationarity is understood in the sense of Clarke, requiring only that for a given target function $\varphi : \mathbb{R}^n \mapsto \mathbb{R}$ the zero vector is contained in the generalized gradient $\partial^C \varphi(x)$ at the stationary point x . Even though the algebraic inclusion $0 \in \partial^C \varphi(x)$ must hold at minima and maxima alike it is usually the only optimality condition found throughout the literature on nonsmooth optimization. The reason is of course that defining, computing, and reasoning about higher order information for general nonsmooth functions is extremely difficult. A more selective unconstrained optimality condition is that $0 \in \partial^M \varphi(x)$ where $\partial^M \varphi(x) \subset \partial^C \varphi(x)$ represents the generalized gradient of Mordukhovich [14]. However, this generally nonconvex set is usually even more difficult to determine and represent than the Clarke gradient.

For piecewise linear problems, Huang et al. in [8] consider a so-called compact representation of the form

$$\varphi(x) = \sum_{i=1}^M \beta_m \max\{l_1^m(x), \dots, l_n^m(x)\} \quad (3)$$

with $\beta_m = \pm 1$ and $l_i^m(x) = a_{mi}^\top x + b_{mi}$ affine functions. They derive a first order necessary and sufficient first order condition based on active gradients a_{mi} of the affine functions involving also some kind of multipliers. Furthermore, for computing a ℓ_1 solution to an overdetermined linear system, in [2] the authors used a reformulation as linear programming problem for the explicit construction of a descent direction. In [1] this approach was extended to incorporate also linear constraints.

To extend the available optimality conditions for nonsmooth target functions, in the present paper we derive constructive optimality conditions for a large class of piecewise smooth functions that covers the scenarios cited above and more general situations.

The first and second Nesterov variants are piecewise smooth and the second one is even piecewise linear. Therefore the optimality conditions presented in this paper apply and since our LIKQ holds in either case globally they distinguish nicely between minima and saddle points. Moreover, at the latter they constructively yield a descent direction, and that a little surprising without any combinatorial effort whatsoever. Naturally, the two chained Nesterov variants are just used for illustrative purposes and our framework applies to the much larger class of problems defined below. While the Mordukhovich criterion $0 \in \partial^M \varphi(x)$ works well on the Nesterov variants in that it does not accept non minimal stationary points we will see on an extremely simple example that it need not be sufficient even on a univariate piecewise linear function.

Provided the directional derivative

$$\varphi'(x; \Delta x) \equiv \lim_{t \searrow 0} \frac{1}{t} [\varphi(x + t\Delta x) - \varphi(x)]$$

exists for any $\Delta x \in \mathbb{R}^n$, it must of course be nonnegative for x to be a local minimizer of φ . In other words, local minimality at x requires

$$\varphi'(x; \Delta x) \geq 0 \quad \text{for all } \Delta x \in \mathbb{R}^n \tag{4}$$

but we are reluctant to call this rather complex assertion an *optimality condition*. The last term conjures up a KKT-like finite system of equalities and inequalities, whose validity at a candidate minimizer point is easier to check than verifying the local minimality of φ at x more or less directly. All we have gained here is a reduction by one degree of freedom, since now we have to demonstrate that the positively homogeneous function $\varphi'(x; \Delta x)$ is nonnegative for Δx in some $n - 1$ dimensional unit sphere in \mathbb{R}^n . A priori, that is not much simpler than directly checking that $\varphi(\tilde{x}) - \varphi(x)$ is nonnegative for \tilde{x} in some arbitrarily small n dimensional ball about x . So we can only gain substantially when $\varphi'(x; \Delta x)$ is in some sense simpler as a function of Δx than the underlying $\varphi(x)$ is as a function of x . In the smooth case $\varphi'(x; \Delta x) \equiv \nabla \varphi(x)^\top \Delta x$ is linear and (4) is clearly equivalent to the stationarity condition $\nabla \varphi(x) = 0$.

Generally, locally Lipschitz continuous functions need not be directionally differentiable at all, but the counter examples are somewhat contrived from an applications point of view. When $\varphi(x)$ is piecewise smooth it follows as in [16] that $\varphi'(x; \Delta x)$ is piecewise linear, continuous, and positively homogeneous with respect to $\Delta x \in \mathbb{R}^n$. Hence the condition (4) simply says that the piecewise linearization of φ at x must have a local minimizer at $\Delta x = 0$. Consequently, we will then say that x is a *first order minimal* or *linearized minimal* point of φ . Interestingly, it was found in [7] on the second Nesterov variant that it may take hundreds of thousands of random tries to find a descent direction from stationary points that violate (4). This means of course that the conical sector of directions

along which descent is possible has a comparatively small volume on the unit sphere. Nevertheless, provided we have a suitable representation it is clear that the condition (4) should be checkable by a finite deterministic procedure. Probably, the main contribution of this paper is a simple constructive procedure to check whether (4) holds and failing this to produce a descent direction.

The paper is organized as follows. In the remainder of this first section we lay out the framework of piecewise smooth objectives. In Section 2, we introduce the important LIKQ concept. In the following, central Section 3 we derive first and second order optimality conditions for the *localized* case where all kinks are active at the prospective minimizer point. We obtain both necessary and sufficient conditions and also derive descent directions and parabolas if the latter are violated. In Section 4 we perform the reduction from a general to a localized problem, which yields the optimality conditions in general form as our main result in Proposition 4. The paper concludes with a short summary and outlook.

Objectives in abs-normal form

We consider the class of objective functions that exhibit so-called Level-1 non-smoothness in that they are defined as compositions of smooth elemental functions and the absolute value function $\text{abs}(x) = |x|$. Hence they may also include $\max(x, y)$, $\min(x, y)$, and the positive part function $\max(0, x)$, which can be easily cast in terms of the absolute value, using for example

$$\max\{x, y\} = \frac{1}{2}(x + y + |x - y|).$$

The inclusion of the Euclidean norm as an elementary function leads to Level-2 objectives that are still Lipschitz continuous and lexicographically differentiable [15] but no longer piecewise smooth. For that more general class of problems the present analysis does not appear to be easily adapted.

By successively numbering all arguments of absolute value evaluations as *switching variables* z_i for $i = 1 \dots s$ we obtain a piecewise smooth system in abs-normal form [3]. Specifically with $F : \mathbb{R}^{n+s} \mapsto \mathbb{R}^s$ and $f : \mathbb{R}^{n+s} \mapsto \mathbb{R}$ at least twice differentiable in the region of interest we have

$$z = F(x, |z|), \tag{5}$$

$$y = f(x, |z|). \tag{6}$$

For example, the first Nesterov variant $\varphi_1(x)$ can be rewritten with $s = n - 1$ as

$$z_i = F_i(x, |z|) \equiv x_{i+1} - 2x_i^2 + 1 \quad \text{for } i = 1 \dots s$$

with the scalar-valued objective f given by

$$f(x, |z|) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} |z_i|.$$

Also, the problems considered by Huang et al. in [8] and Conn et al. in [2, 1] fit into this setting where in the latter case linear equations for example are used to eliminate certain constraints or their violation is added as ℓ_1 penalty.

For the abs-normal form, $L \equiv F_{|z|} \equiv \partial F(x, |z|) / \partial |z|$ is strictly lower triangular so that one obtains the components of $z = z(x)$ one by one as piecewise smooth Lipschitz continuous functions of x . Therefore we sometimes write

$$\varphi(x) \equiv f(x, |z(x)|)$$

to denote the objective directly in terms of the independent variable vector x . Throughout the paper the symbol $|z|$ is sometimes viewed as a (nonnegative) variable vector in its own right. Accordingly, the other partial derivatives of $F(x, |z|)$ and $f(x, |z|)$ will be denoted by

$$Z \equiv \frac{\partial}{\partial x} F(x, |z|) \in \mathbb{R}^{s \times n}, \quad a = \frac{\partial}{\partial x} f(x, |z|) \in \mathbb{R}^n, \quad b = \frac{\partial}{\partial |z|} f(x, |z|) \in \mathbb{R}^s.$$

The structural nilpotency degree of L , i.e., the smallest number $\nu \leq s$ such that $L^\nu = 0$, is called the *switching depth* of f in the given representation. Problems are called smooth with $\nu = 0$ when $s = 0$ and $L = \emptyset$, and simply switched with $\nu = 1$ when $L = 0 \in \mathbb{R}^{s \times s}$ for some $s > 0$. Stating the compact representation (3) considered by Huang et al. in abs-normal form, one finds that it has switching depth $\log_2 M$. The number ν can also be interpreted as the maximal number of absolute value nodes on a path connecting an independent variable to a dependent variable in the dependency graph representing the evaluation procedure of $\varphi(x)$ (see [5]).

In other words ν quantifies how often nonsmooth elementals are composed on each other, which represents a difficult challenge for any kind of chain-rule. That difficulty does not arise in KKT conditions for smooth nonlinear optimization problems (NLOPs) and other classical complementarity systems, as they are usually simply switched. In that case $z = z(x) = F(x)$ is a smooth vector function and only $f(x, |z(x)|)$ or a corresponding vector function are nonsmooth. The same comparatively simple situation arises when one wishes to minimize a nonsmooth norm $\|F(x)\|_p$ with $p \in \{1, \infty\}$ of a smooth residual vector $F : \mathbb{R}^n \mapsto \mathbb{R}^n$.

It was shown in [4] for simply switched piecewise linear systems of n equations in as many variables that the intermediate solvability status *open but not bijective* (see [16]) cannot occur. We view this as an indication that the full richness of piecewise linear systems arises only on multiply switched systems where the kink surfaces are not just smooth hypersurfaces.

The combinatorial aspect of the evaluation can be expressed in terms of the vector $\sigma(x) \equiv \mathbf{sgn}(z(x))$ and the corresponding diagonal matrix $\Sigma = \mathbf{diag}(\sigma)$. Especially in view of the \mathcal{VU} decomposition [12] we will sometimes assume the following natural property:

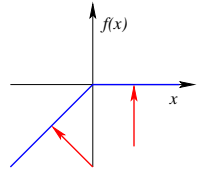
Definition 1 (Nonredundancy). *The abs-normal representation of the objective function $\varphi(x)$ is called nonredundant at the point x if the sparsity patterns of $L = L(x)$ and $b = b(x)$ and their values are such that for at least one signature matrix $\Sigma = \text{diag}(\sigma)$ with $\sigma \in \{-1, 1\}^s$ we have componentwise*

$$|(\Sigma - L^\top)^{-1}b| > 0. \quad (7)$$

If this holds for at least one argument x the abs-normal form itself is called nonredundant.

Nonredundancy requires that no row of the bordered matrix $[L^\top(x), b(x)] \in \mathbb{R}^{s \times (s+1)}$ vanishes identically, so that each switching variable $z_j(x)$ influences the objective $y(x)$ either directly via $b_j(x) \neq 0$ or indirectly through switching variables $z_i(x)$ with $i > j$. Any $z_j(x)$ for which this is not the case can simply be dropped from the computational graph because it does not impact the objective at all. Moreover, at a point of nonredundancy x the corresponding entries of $L(x)$ and $b(x)$ must be generic enough such that for at least one definite Σ the vector in Eq. (7) does not have a zero component due to cancellation. An extreme case of redundancy would be $b = 0$ constantly in which case the objective $\varphi(x) = f(x, 0)$ is just a smooth function. A nonredundant example would be $b = e_s \in \mathbb{R}^s$ the last Cartesian basis vector and $L(x)$ having a non-vanishing subdiagonal. When F and f are real analytic the set of exceptional points where a generally nonredundant system is redundant is a semi-algebraic set and has therefore measure zero. Everywhere else Lemma 4 yields an explicit representation of \mathcal{V} as tangent space of the affine hull of the generalized gradient at x .

An extremely simple example of Level-1 nonsmoothness is given by



$$\begin{aligned} \varphi(x) &= \min(x, 0) = \frac{1}{2}(x - |x|) \\ &= f(x, |z|) \equiv \frac{1}{2}(x - |z|) \quad (8) \\ &\text{with } z = F(x) \equiv x. \end{aligned}$$

Here we have only one independent variable which is also the only switching variable so that $Z = 1, L = 0, a = \frac{1}{2} = -b$, and the switching depth is $\nu = 1$. Obviously this function does not have a minimum, but at $x = 0$ the trivial derivative 0 belongs not only to the Clarke generalized gradient $\partial^C \varphi(0) = [-1, 0]$ but also to the Mordukhovich generalized gradient $\partial^M \varphi(0) = \{-1, 0\}$. The latter is here just the limiting gradient, i.e., the set of all limits of proper gradients in the vicinity and the former is its convex hull. Neither gives an indication that the origin is far from being a minimum, since -1 is a direction of linear descent.

Details for the Nesterov variants

The zeroth Nesterov variant $\varphi_0(x)$ is smooth with Z, L , and b empty as there are no switching variables at all. Thus we have $f(x, |z|) = \varphi_0(x)$ so that simply

$a = \nabla_x \varphi_0(x)$. For the first Nesterov variant $\varphi_1(x)$ at any reference point x the matrices describing the abs-normal form are given by

$$Z(x) = \begin{bmatrix} -4x_1 & 1 & 0 & \dots\dots\dots & 0 & 0 \\ 0 & -4x_2 & 1 & \dots\dots\dots & 0 & 0 \\ 0 & 0 & -4x_3 & \dots\dots\dots & 0 & 0 \\ \dots & \dots & \dots & \dots\dots\dots & \dots & \dots \\ 0 & 0 & 0 & \dots\dots\dots & 1 & 0 \\ 0 & 0 & 0 & \dots\dots\dots & -4x_s & 1 \end{bmatrix} \in \mathbb{R}^{(n-1) \times n} \quad (9)$$

and

$$L = 0 \in \mathbb{R}^{(n-1) \times (n-1)}, \quad a(x) = (\tfrac{1}{2}(x_1 - 1), 0 \dots 0) \in \mathbb{R}^n, \quad b = (1 \dots 1) \in \mathbb{R}^{n-1}.$$

Since $L = L^1 = 0$ without being empty, i.e., $s = n - 1 > 0$ for $n > 1$ as we will assume, the switching depth ν of this problem is 1, i.e., this example is simply switched. Finally, we note that the representation is nonredundant everywhere since b is dense and constant.

For the second version we have $s = 2n - 1$ switching variables, namely

$$z_i = F_i(x, |z|) = x_i \quad \text{for } 1 \leq i < n, \quad z_n = F_n(x, |z|) = x_1 - 1,$$

and

$$z_{n+i} = F_{n+i}(x, |z|) = x_{i+1} - 2|z_i| + 1 \quad \text{for } 1 \leq i < n.$$

Hence we obtain the matrices and vectors

$$Z = \begin{bmatrix} I_{n-1} & 0 \\ I_{n-1} & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{s \times n}, \quad L = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -2I_{n-1} & 0 & 0 \end{bmatrix} \in \mathbb{R}^{s \times s},$$

$$a = 0 \in \mathbb{R}^n, \quad b^\top = (0, 0 \dots 0, \tfrac{1}{4}, 1 \dots 1) \in \mathbb{R}^{(n-1)+n}.$$

Here and throughout I_k denotes the identity matrix of dimension k and the symbol 0 pads with zeros to achieve the specified dimensions.

To establish nonredundancy we note that for $\Sigma = I_s$ we have $b^\top(\Sigma - L)^{-1} = (e, \frac{1}{4}, -e)$ with e denoting a vector of ones, which is again dense everywhere. For this second Nesterov version we have $L \neq 0 = L^2$ which means the switching depth ν is now 2. Also note that the derivative matrices and vectors are all constant since φ_2 itself is piecewise linear.

Superposition of nonsmoothness

The next example shows that the abs-normal representation covers the repeated superposition of nonsmoothness. To the best knowledge of the authors, this situation has rarely been considered in nonsmooth analysis.

For a given $\varepsilon \in \mathbb{R}$, the target function $\varphi : \mathbb{R}^n \mapsto \mathbb{R}$ is defined by

$$\varphi(x) = y_n + \varepsilon \sum_{i=1}^{n-1} y_i \quad \text{with} \quad y_i = |x_i - y_{i-1}|, \quad i = 1, \dots, n, \quad \text{and} \quad y_0 = 0.$$

We rewrite $\varphi(x)$ in abs-normal form as

$$f(x, |z|) = |z_n| + \varepsilon \sum_{i=1}^{n-1} |z_i|, \quad z_1 = x_1, \quad \text{and} \quad z_i = x_i - |z_{i-1}|, \quad i = 2, \dots, n. \quad (10)$$

Then, one has $Z = I$ and $L = \mathbf{subdiag}(-1)$, which results in a switching depth $\nu = n$, i.e., the nonsmoothness is maximally superimposed. In particular that means that only the first kink surface $z_1 = 0$ is smooth, already $z_2 = 0 = x_2 - |x_1|$ has the two normals $(1, \pm 1, 0, \dots, 0)$ and the kink surface $z_k = 0$ has 2^{k-1} different normals. Correspondingly, there are 2^n different subgradients in the vicinity of the optimal point $x^* = 0 \in \mathbb{R}^n$ that would all contribute to the Clarke gradient at x^* . Furthermore, one obtains $a = 0 \in \mathbb{R}^n$ and $b = (\varepsilon, \dots, \varepsilon, 1)^\top \in \mathbb{R}^n$. For reasons illustrated later, we will call this target function the gradient cube example.

For $\varepsilon > 0$, the function φ has an isolate minimum at $x = 0 \in \mathbb{R}^n$. For $\varepsilon = 0$, the minimizer is degenerate and for $\varepsilon < 0$, the point $x = 0 \in \mathbb{R}^n$ is a saddle point of φ as shown in Fig. 1 for $n = 2$. Here, one can also see that φ is in this case not convex provided $0 < \varepsilon < 1$. Hence, this is an example for a piecewise linear function with an isolated minimizer in whose immediate neighborhood the function is not convex.

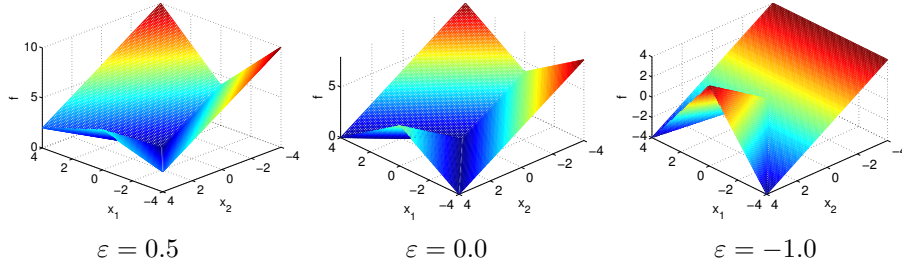


Figure 1: Gradient cube problem for $n = 2$ and different values of ε

2 Linear kink independence

The signature vectors define the domains

$$P_\sigma = \{x \in \mathbb{R}^n \mid \mathbf{sgn}(z(x)) = \sigma\}.$$

Here the question arises when and where these signature domains P_σ look essentially like polyhedra. More specifically, one may ask when they are smooth manifolds of dimension $|\sigma| \equiv \|\sigma\|_1$ defined by the active switch set

$$\alpha = \alpha(x) \equiv \{1 \leq i \leq s \mid \sigma_i(x) = 0\} \quad \text{of size} \quad |\alpha(x)| = s - |\sigma(x)|.$$

It follows by continuity of F that P_σ must be open but possibly empty if σ is *definite* in that all its components are nonzero. Generally we have for any nonempty P_σ

$$\dim(P_\sigma) \geq n + \|\sigma\|_1 - s = n - s + \sum_{i=1}^s |\sigma_i|.$$

When equality holds we call the signature σ *nondegenerate* and otherwise *critical*. In particular 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 [16, Chapter 2].

For each fixed σ and thus Σ the system

$$z = F(x, \Sigma z)$$

is smooth and has by the implicit function theorem a locally unique solution $z^\sigma = z^\sigma(x)$ with a well defined Jacobian

$$\nabla z^\sigma \equiv \frac{\partial}{\partial x} z^\sigma = (I - L\Sigma)^{-1} Z \in \mathbb{R}^{s \times n}, \quad (11)$$

where Z and L are evaluated at $(x, z^\sigma(x))$. Of particular interest is the submatrix

$$\nabla z_\alpha^\sigma(x) \equiv (e_i^\top \nabla z^\sigma(x))_{i \in \alpha} \in \mathbb{R}^{|\alpha| \times n}.$$

Definition 2 (LIKQ). *We say that the linear independence kink qualification is satisfied at a point $x \in \mathbb{R}^n$ if for $\sigma = \sigma(x)$ the active Jacobian*

$$V(x)^\top \equiv \nabla z_\alpha^\sigma(x) \equiv (e_i^\top \nabla z^\sigma(x))_{i \in \alpha} \in \mathbb{R}^{|\alpha| \times n} \quad (12)$$

has full row rank $|\alpha|$, which requires in particular that $|\sigma| \geq s - n$ so that $V(x)$ is a short, fat matrix. Moreover, there then exists an orthogonal matrix $U(x) \in \mathbb{R}^{n \times (n - |\alpha|)}$ such that

$$V(x)^\top U(x) = 0 \quad \text{and} \quad U(x)^\top U(x) = I.$$

Wherever LIKQ holds the function φ will be smooth within P_σ but have kinks along certain normal directions. This geometry is similar to the one described by means of a $\mathcal{V}\mathcal{U}$ decomposition by Mifflin and Sagastizábal [13] and Lewis [10]. Here we can define at a point x the pair of orthogonal subspaces

$$\mathcal{V}(x) \equiv \text{range}(V(x)) \quad \text{and} \quad \mathcal{U}(x) \equiv \mathcal{V}(x)^\perp = \text{range}(U(x)).$$

We will see in the following section that under the nonredundancy assumption (7) the subspace $\mathcal{V}(x)$ is indeed the tangent space of the affine hull of the generalized gradient at x . Without the additional assumption, for example when $b(x) = 0$, the tangent space can be larger than the range of the transposed Jacobian matrix $V(x)$. Correspondingly, the orthogonal complement $\mathcal{U}(x)$ may turn out smaller than in the classical definition.

For piecewise linear systems, whose polyhedral decomposition is completely determined by the matrices Z, L , and the vector $c = F(0, 0)$ it is generic that all polyhedra satisfy the linear independence condition so that

$$\{x \in B \mid \text{sgn}(z(x)) = \sigma\} = \{x \in B \mid z_i^\sigma(x) = 0 \text{ for } i \in \alpha\}$$

where B is a suitably small neighborhood of a given point.

When the problem is simply switched LIKQ is equivalent to LICQ for the constrained set

$$M_\alpha \equiv \{x \in \mathbb{R}^n \mid z_i(x) = 0 \text{ for } i \in \alpha\}.$$

Then each one of the $z_i(x)$ is a differentiable function and the solution sheets $\{x \in \mathbb{R}^n : z_i(x) = 0\}$ for $i \in \alpha$ are locally smooth hypersurfaces. Then LIKQ requires the active Jacobian $V(x)$ to have full rank making M_α a smooth manifold of co-dimension $|\alpha|$. In that sense LIKQ is merely a generalization of LICQ. However, we note that all points in a neighborhood of x are feasible arguments of our nonsmooth problem so that there is no proper constraint active at x . On the other hand there may be plenty of kinks, which justifies our terminology LIKQ.

In general there is no reason why LIKQ should be violated and locally it can always be achieved by an arbitrary small perturbation of Z . To see this we observe that the determinant of any square submatrix of ∇z^σ as defined in (11) is a polynomial in the entries of Z , which does not vanish if we place a (permuted) identity in the corresponding position of Z and set all other entries to zero. Hence each square submatrix of ∇z^σ is nonsingular for almost all small perturbations of Z and also L . On the simple example (8) described above $x = z = 0$ is the only kink and we have $Z = 1 \in \mathbb{R}$ so that LIKQ certainly holds. Now, we demonstrate its validity on the first and second Nesterov examples. It holds trivially for the zeroth, which has no kinks at all.

Lemma 1. (*LIKQ for Nesterov*) *The functions defined in (1) and (2) with their abs-normal forms given above satisfy LIKQ globally, i.e., throughout \mathbb{R}^n .*

Proof. For (1) we have $L = 0$ so that for all $\sigma \in \{-1, 0, 1\}^s$ identically $\nabla z^\sigma = Z$. Since Z takes the form (9) it has at all points x linearly independent rows and any subset of them forms also a Jacobian of full rank. Hence LIKQ is satisfied everywhere.

For the second variant (2), we get with $\Sigma_{n-1} \equiv \text{diag}(\sigma_i)_{i=1 \dots n-1}$

$$\begin{aligned} \nabla z^\sigma &= \begin{bmatrix} I_{n-1} & 0 & 0 \\ 0 & 1 & 0 \\ -2\Sigma_{n-1} & 0 & I_{n-1} \end{bmatrix}^{-1} \begin{bmatrix} I_{n-1} & 0 \\ I_{n-1} & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} I_{n-1} & 0 & 0 \\ 0 & 1 & 0 \\ 2\Sigma_{n-1} & 0 & I_{n-1} \end{bmatrix} \begin{bmatrix} I_{n-1} & 0 \\ I_{n-1} & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} 1 & 0 & 0 & \dots\dots\dots & 0 & 0 \\ 0 & 1 & 0 & \dots\dots\dots & 0 & 0 \\ 0 & 0 & 1 & \dots\dots\dots & 0 & 0 \\ \dots & \dots & \dots & \dots\dots\dots & \dots & \dots \\ 0 & 0 & 0 & \dots\dots\dots & 0 & 0 \\ 0 & 0 & 0 & \dots\dots\dots & 1 & 0 \\ 1 & 0 & 0 & \dots\dots\dots & 0 & 0 \\ 2\sigma_1 & 1 & 0 & \dots\dots\dots & 0 & 0 \\ 0 & 2\sigma_2 & 1 & \dots\dots\dots & 0 & 0 \\ \dots & \dots & \dots & \dots\dots\dots & \dots & \dots \\ 0 & 0 & 0 & \dots\dots\dots & 1 & 0 \\ 0 & 0 & 0 & \dots\dots\dots & 2\sigma_{n-1} & 1 \end{bmatrix} \in \mathbb{R}^{s \times n},$$

where $s = n - 1 + 1 + n - 1 = 2n - 1$. Now we have to show that no matter what the activity $\alpha = \alpha(x) \subset \{1, \dots, 2, n - 1\}$ at any point $x \in \mathbb{R}^n$ is we always get a Jacobian of full rank.

Let us first consider the *linking* switches

$$\sigma_{n+i} = \mathbf{sgn}(z_{n+i}) = \mathbf{sgn}(x_{i+1} - 2|x_i| + 1) \quad \text{for } i = 1 \dots n - 1.$$

If any one of them is inactive there is locally no connection between the values of the variables x_k for $k \leq i$ and the variables x_k for $k > i$ at all. Consequently the corresponding rows of the Jacobian ∇z^σ do not share nonzeros in any column and we can consider them separately. The remaining problem to the left will have exactly the same structure as the full one with n reduced to i and the one on the right has also the same structure except that the first variable x_{i+1} does only occur in two switches rather than three as is true for x_1 . That only simplifies the switching structure and thus makes the satisfaction of LIKQ even more likely. Hence we may assume without loss of generality that all linking switches are active. As observed in [7] that means either that all x_i are nonzero or exactly one x_i is zero. In the first case the first $n - 1$ rows of the Jacobian are inactive and only a subset of the bottom n rows can be active. They form a unitary lower triangular and thus nonsingular matrix so that LIKQ must hold. Now if x_1 is zero the first row is active but the n -th not and we have again a unitary lower triangular matrix. If $x_i = 0$ for some $i > 1$, one can eliminate from ∇z^σ the i -th column, the first n rows, and the $(n + i)$ -th row to obtain a $(n - 1) \times (n - 1)$ submatrix that is block diagonal with one upper triangular and one unitary lower triangular block. The diagonal of the upper triangular matrix has the entries $\sigma_j \neq 0$ for $j = 1 \dots i - 1$ so that the submatrix formed by the i -th row and the last $n - 1$ rows is again nonsingular. Hence LIKQ is again satisfied globally. \square

3 Optimality conditions for localized problem

We are looking for necessary and sufficient conditions for x to be a minimizer of $f(x, |z(x)|)$ on some ball B . First we will assume that at the given point x

the vector of switching variables vanishes completely so that $z = z(x) = 0$. We will then say that the switching is *localized* at x . In the next section it will be shown how this *localization* can be achieved by an implicit reduction and how the reduced quantities can be obtained from the original ones.

Optimality on trunk problem

So now we assume $s = |\alpha| \leq n$ with $Z = F_x(x, 0) \in \mathbb{R}^{s \times n}$ attaining full row rank according to the LIKQ condition. For x to be a local minimizer in some neighborhood of \mathbb{R}^n it must firstly be a local minimizer of the equality constrained problem

$$\min f(x, 0) \quad \text{s.t.} \quad 0 = F(x, 0).$$

We will call this the smooth *trunk* problem as all z_i have been clamped at 0 so that no nonsmoothness arises at all. By our LIKQ assumption the constraint Jacobian Z has full rank and its nullspace is spanned by $U \in \mathbb{R}^{n \times (n-s)}$. Thus we obtain with LICQ immediately the classical first order condition

$$a^\top + \lambda^\top Z = 0 \quad \text{with} \quad \lambda \in \mathbb{R}^s. \quad (13)$$

We will refer to this property as *tangential stationarity*. Furthermore, we obtain the second order necessary condition

$$U^\top H U \succeq 0 \quad \text{with} \quad H(x, \lambda) \equiv f(x, 0)_{xx} + (\lambda^\top F(x, 0))_{xx} \in \mathbb{R}^{n \times n}, \quad (14)$$

i.e., positive curvature, which turns into a sufficient optimality condition in combination with tangential stationarity if $\det(U^\top H U) > 0$, i.e., the projected Hessian is nonsingular. We will call this property *trunk optimality*.

On the zeroth, smooth Nesterov example (0) we have $s = 0$ and thus simply the usual unconstrained optimality conditions

$$a = \nabla f(x) = 0 \quad \text{and} \quad H(x) = \nabla^2 f(x) \succeq 0$$

which holds strictly at the unique stationary point $x = (1, \dots, 1) \in \mathbb{R}^n$.

Optimality on branch problems

Of course being minimal on the trunk problem is not sufficient for x to be minimal on an open neighborhood in \mathbb{R}^n . For that it must be for any definite $\sigma \in \{-1, 1\}^s$ and corresponding $\Sigma = \mathbf{diag}(\sigma)$ a local minimizer of the optimization problem

$$\min f(x, \Sigma z) \quad \text{s.t.} \quad z = F(x, \Sigma z) \quad \text{and} \quad \Sigma z \geq 0.$$

We will refer to these smooth, inequality constrained problems as *branch* problems, since they extend the common trunk problem in various directions. For $n = 3$ and $s = 2 = |\alpha|$ the trunk problem and the related branch problems are sketched in Fig. 2.

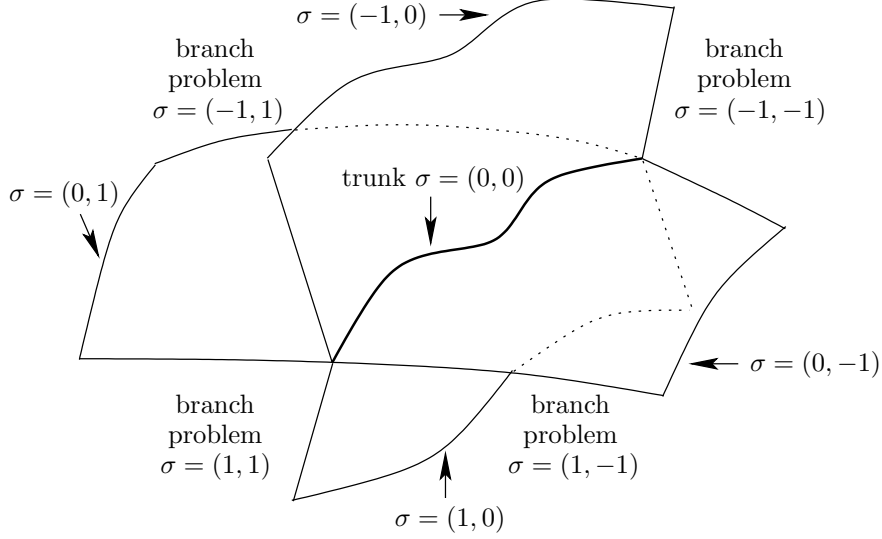


Figure 2: Trunk problem and branch problems for $n = 3$ and $s = 2 = |\alpha|$

Each of the 2^s branch problems fixes one $\Sigma = \mathbf{diag}(\sigma)$ with $\sigma \in \{-1, 1\}^s$. Then, we may rewrite such a problem in terms of $\bar{z} = \Sigma z$ with the given Σ as

$$\min f(x, \bar{z}) \quad \text{s.t.} \quad \Sigma \bar{z} = F(x, \bar{z}) \quad \text{and} \quad \bar{z} \geq 0. \quad (15)$$

The Jacobian of all equality and inequality constraints is given by

$$\begin{bmatrix} Z & L - \Sigma \\ 0 & I_s \end{bmatrix} = \begin{bmatrix} L - \Sigma & 0 \\ 0 & I_s \end{bmatrix} \begin{bmatrix} (L - \Sigma)^{-1} Z & I_s \\ 0 & I_s \end{bmatrix} \in \mathbb{R}^{2s \times (n+s)}. \quad (16)$$

Since by the LIKQ assumption the rows of Z are linearly independent the same is true for the full Jacobian and any of its submatrices obtained when some of the sign conditions on \bar{z} are inactive. Hence, LIKQ for the localized nonsmooth problem implies LICQ for all branch problems. The first n components of the gradient of f and the first n columns of the constraint Jacobians are identical for all branch problems so that the equality constraints have for all branch problems the same vector of Lagrange multipliers λ as the trunk problem. That is quite remarkable and simplifies the situation enormously. In addition to the tangential stationarity condition (13) we get for the branch problems with a second vector of Lagrange multipliers $\mu \in \mathbb{R}^s$ corresponding to the inequality constraints the KKT condition

$$b^\top + \lambda^\top (L - \Sigma) \equiv \mu \geq 0. \quad (17)$$

This inequality must hold for all signature matrices Σ including any definite $\Sigma = \mathbf{diag}(\sigma)$ with $\sigma_i \lambda_i \geq 0$ so that $\Sigma \lambda = |\lambda|$. All such choices generate the strongest condition namely that the Lagrange multipliers μ associated with the

sign conditions $\bar{z} \geq 0$ will be nonnegative or strictly positive if and only if

$$b + L^\top \lambda \geq |\lambda| \quad \text{or} \quad b + L^\top \lambda > |\lambda|, \quad (18)$$

respectively. Obviously, these first order optimality conditions are very easy to check and contrary to what one might have expected there is no combinatorial effort at all. Given λ as defined by tangential stationarity we interpret (18) as a componentwise lower bound on the gradient b of $f(x, |z|)$ with respect to $|z|$. Hence, we will call this inequality the *normal growth* condition. Figure 3 illustrates the relations between the optimality conditions derived so far.

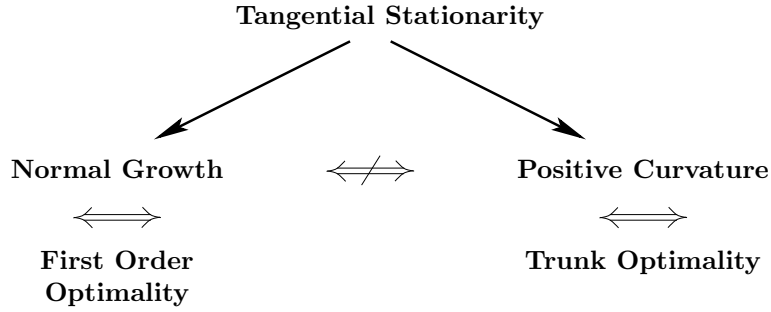


Figure 3: Relation of optimality conditions

When all components of λ are nonzero, among all branch problems there is a unique one with $\Sigma = \mathbf{diag}(\mathbf{sgn}(\lambda))$ such that all Lagrange multipliers μ are minimal at its stationary point x . This critical branch is a promising choice for a successive descent algorithm.

Linear first order sufficiency and generalized stationarity

Now suppose that both $F(x, |z|)$ and $f(x, |z|)$ are linear, or more precisely affine functions so that we have a piecewise linear objective function $\varphi(x)$. Then the branch problems given in (15) are all linear optimization problems and the weak form of (18) derived from (17) is already sufficient for optimality. The same is true in the nonlinear case when $|\alpha| = n$ kinks are active so that the second order condition is trivial as $\tilde{H} = U^\top H U$ is empty. Hence we conclude:

Lemma 2 (First order sufficiency). *When F and f are linear or $|\alpha| = s = n$, a point x where LIKQ holds is a local minimizer if and only if*

$$a^\top + \lambda^\top Z = 0 \quad \text{and} \quad b^\top \geq |\lambda|^\top - \lambda^\top L. \quad (19)$$

Moreover, the second condition is equivalent to

$$b^\top + \lambda^\top L \geq \lambda^\top \Sigma \quad \text{and} \quad b^\top + \lambda^\top L \geq \lambda^\top (-\Sigma)$$

for any particular $\Sigma = \mathbf{diag}(\sigma)$ with $\sigma \in \{-1, 1\}^s$.

The last assertion is trivial to check and allows the following interpretation. Suppose any one of the branch problems defined by a given Σ has the minimizer x , which could for example be located by using the simplex method. Then we can flip all signs of the σ_i around and consider the *reflected* branch problem defined by $-\Sigma$. If it also has the common *pivot* point x as minimizer, the second condition of Eq. (19) in Lemma 2 must be satisfied and x must be a minimizer for the original nonsmooth problem itself. Otherwise linear optimization applied to the reflected problem will lead to a lower local minimum of the reflected problem if that is bounded at all. One would expect that such a successive search through distinct polyhedra would generate a path of reasonable length before termination, but exponential searches such as the Simplex algorithm performs on the Klee-Minty cube are certainly possible.

For the gradient cube example (10) with $\varepsilon \geq 0$ all kinks are active at the optimal point $x^* = 0 \in \mathbb{R}^n$. Hence, the optimization problem is localized at x^* and from $Z = I$ follows that LIKQ is fulfilled. Furthermore, combining $a = 0 \in \mathbb{R}^n$ and $b = (\varepsilon, \dots, \varepsilon, 1)^\top \in \mathbb{R}^n$ with Eq. (19), one obtains that the first order sufficiency conditions are satisfied for the Lagrange multiplier $\lambda = 0 \in \mathbb{R}^n$ if and only if $\varepsilon \geq 0$. Then the reflection algorithm will reach a minimizer in one step from anywhere and then verify it in a second step. Otherwise, if $\varepsilon < 0$ it will detect unboundedness in the first or second step. Here each *step* requires solving an LP, which might be quite costly. In contrast it was found that established nonsmooth optimization algorithms need hundreds or thousands of iterations to reach the minimum for moderate dimensions and fail completely for $n \gg 50$.

The sketched strategy can also be applied to nonlocalized and nonlinear problems. Then one may apply a reflection algorithm that solves a sequence of NLOPs always flipping the active switching variables at the computed local minimizers until the process becomes stationary due to the reflected problem also having the common pivot as a minimizer. The *successive reflection* method has been applied to the second Nesterov variant and does indeed always yield finite convergence. For $n = 2$ the situation is depicted in Fig. 4, where the two vertical, straight kinks at $x_1 = 0$ and $x_1 = 1$ and a third hooked kink where $x_2 = 2|x_1| - 1$ are drawn as blue lines. There are six open polyhedra with the signatures $\sigma \in \{(-1, -1, -1), (-1, -1, 1), (1, -1, -1), (1, -1, 1), (1, 1, -1), (1, 1, 1)\}$ because the signatures $(-1, 1, -1)$ and $(-1, 1, 1)$ cannot arise as x_1 cannot be smaller than 0 and larger than 1 at the same time. The global minimum is attained at $x_* = (1, 1)$ but there is another stationary point at $(0, -1)$ both of which are marked with a green circle.

Suppose we begin the minimization in the polyhedron $(-1, -1, -1)$ leading to its minimizer at the stationary point $(0, -1)$. There the first and the third kink are active so that reflection leads to the polyhedron $(1, -1, 1)$. On that polyhedron $(0, -1)$ is not a minimizer as the minimum of that branch problem is attained at $(1, 1)$. Reflection at that pivot leads into the polyhedron $(1, 1, -1)$ which also has the pivot as its minimizer so the process stops as we have found a local minimizer, which here happens to be global as well. This optimization process is also illustrated in Fig. 4.

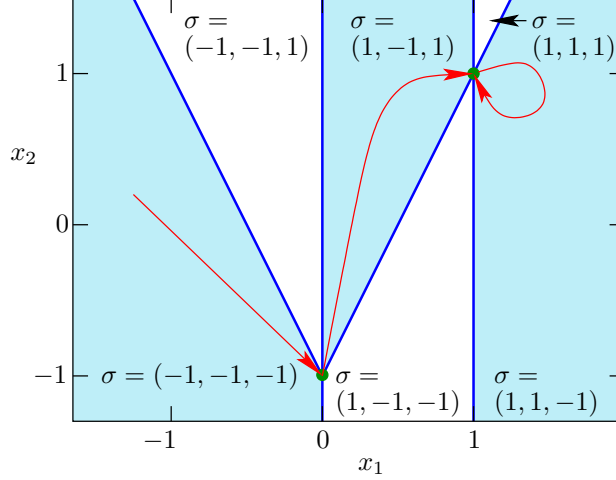


Figure 4: Successive reflection method for Nesterov variant (2) for $n = 2$

In general the basic reflection strategy can break down wherever LIKQ is violated so that some modifications for handling degeneracy must be developed. Almost as a corollary of the last lemma we obtain the following result.

Lemma 3 (Relation to stationarity concepts). *Under the assumption of LIKQ, tangential stationarity and normal growth are equivalent to first order minimality. The latter implies always Clarke stationarity, which in turn under LIKQ implies tangential stationarity but not necessarily normal growth. Mordukhovich stationarity is stronger than Clarke stationarity but in certain cases still weaker than first order minimality.*

Proof. Since Clarke and Mordukhovich stationarity are necessary conditions for optimality, they must be implied by first order minimality, which is already sufficient for minimality in the piecewise linear case. To prove the second assertion we note that as observed in [6] also in the nonlinear case, the limiting gradients in the vicinity of x are given by

$$g_\sigma^\top \equiv a^\top + b^\top(\Sigma - L)^{-1}Z \quad \text{for } \sigma \in \{-1, 1\}^s. \quad (20)$$

Numbering the 2^s possibilities of σ as σ_j and the corresponding diagonal matrices as Σ_j we note that Clarke stationarity is equivalent to the existence of a convex combination

$$0 = \sum_j \gamma_j [a^\top + b^\top(\Sigma_j - L)^{-1}Z] = a^\top + \lambda^\top Z$$

with $\lambda^\top \equiv \sum_j \gamma_j b^\top(\Sigma_j - L)^{-1}$. Hence, Clarke stationarity implies tangential stationarity as claimed. The remaining assertions follow from our trivial exam-

ple (8). At $x = 0$, we have $|\alpha| = 1 = s \leq n$ and the stationarity condition yields $\frac{1}{2} = a = -\lambda Z = -\lambda \cdot 1$ so that $\lambda = -\frac{1}{2}$. Consequently the normal growth condition

$$b + L\lambda = -\frac{1}{2} - 0 \cdot \frac{1}{2} = -\frac{1}{2} \geq |-\frac{1}{2}| = \frac{1}{2}$$

is violated. Hence we find that $x = 0$ can indeed not be a minimizer, whereas the Clarke and Mordukhovich criterion leave that possibility open. That the latter is stronger than the former follows from the second Nesterov example. \square

For the target function defined in Eq. (10) and $\varepsilon = 0$, one has $b = e_n$, i.e., the n -th unit vector. Since all 2^n definite $\sigma \in \{-1, 1\}^n$ appear near the origin, one obtains from Eq. (20) that the limiting gradients in the vicinity of $x^* = 0 \in \mathbb{R}^n$ given by

$$g_\sigma = (\Sigma - L)^{-1}b \quad \text{for } \sigma \in \{-1, 1\}^n$$

are exactly the corners of the unit square in \mathbb{R}^n . For this reason, we call this example the gradient cube example. The resulting Clarke gradient has 2^n corners making it very difficult to deal with it explicitly. Since everything is continuous, a small positive or negative ε will only slightly perturb the corners of the unit cube.

Lemma 3 and the following Lemma 4 also hold in the general nonlocalized case, where not all switching variables vanish since these assertions are invariant with respect to the smooth reduction process discussed in Section 4.

Formula (20) can be used to explicitly characterize the subspace \mathcal{V} and thus the \mathcal{VU} decomposition in the nonredundant case.

Lemma 4 (\mathcal{VU} decomposition in general LIKQ and nonredundant case).

Assuming LIKQ in the localized scenario we have for $V(x)^\top \equiv Z(x)$ and $e = (1, 1, \dots, 1)$, that

$$\text{span}(g_\sigma(x) - g_e(x))_{\sigma \in \{-1, 1\}^s} \subset \mathcal{V}(x) \equiv \text{range}(Z(x)^\top).$$

Moreover, in the nonredundant case where (7) holds for some Σ we have equality of the two subspaces.

Proof. Let us drop the explicit reference to x and consider the vectors $b_\sigma^\top \equiv b^\top(\Sigma - L)^{-\top} \in \mathbb{R}^s$ with σ ranging again over all 2^s definite signature vectors of length s . Then we have $g_\sigma = a + Z^\top b_\sigma$ and see immediately that

$$g_\sigma - g_e = Z^\top(b_\sigma - b_e) \in \text{range}(Z^\top),$$

which proves the first assertion. The assumption for the second assertion certainly requires $b \neq 0$ and thus $b_\sigma \neq 0$ for all σ . Now, if the span of $\{b_\sigma - b_e \mid \sigma \in \{-1, 1\}^s\}$ is the whole \mathbb{R}^s it is clear that the span of the $g_\sigma - g_e$ is the range of Z^\top . The only possibility for this not to be the case is that there exists a nonzero vector $v \in \mathbb{R}^s$ such that

$$0 = (b_\sigma - b_e)^\top v = b^\top(\Sigma - L)^{-1}v - b_e^\top v \quad \text{for all } \sigma \in \{-1, 1\}^s. \quad (21)$$

With $\beta = b_e^\top v$ this is equivalent to the statement that

$$\det C(\sigma) = 0 \quad \text{for} \quad C(\sigma) \equiv \begin{bmatrix} \Sigma - L & v \\ b^\top & \beta \end{bmatrix} \quad \text{and all} \quad \sigma \in \{-1, 1\}^s.$$

Now, if we replace the first diagonal element σ_1 by a real number d_1 we find that because of its multilinearity the determinant of $C(d_1, \sigma_2, \dots, \sigma_s)$ is an affine function of d_1 and thus because it vanishes for $d_1 = \pm 1$ it must be zero throughout. This argument can be applied for each diagonal element so that we get

$$\det \begin{bmatrix} D - L & v \\ b^\top & \beta \end{bmatrix} = 0 \quad \text{for} \quad D = \mathbf{diag}(d) \quad \text{with} \quad d \in \mathbb{R}^s.$$

Then, if we set $D = \delta I_s$ we see that the determinant is a vanishing polynomial in δ whose highest power δ^s has the coefficient β , which must therefore vanish. Combining $0 = \beta = b_e^\top v$ with Eq. (21), we find that

$$0 = b^\top (D - L)^{-1} v \quad \text{with} \quad D = \mathbf{diag}(d) \quad \text{and} \quad \det(D) \neq 0.$$

Differentiating this identity with respect to d_i for $i = 1, \dots, n$, we find that also

$$b^\top (D - L)^{-1} e_i e_i^\top (D - L)^{-1} v = 0 \quad \text{for} \quad i = 1 \dots n.$$

This must hold in particular for $D = \Sigma$ where all components of $b^\top (\Sigma - L)^{-1}$ are nonzero, i.e., in the nonredundant case. Therefore, all components of $(\Sigma - L)^{-1} v$ must be zero and hence we have necessarily $v = 0$. Thus the $b_\sigma - b_e$, $\sigma \in \{-1, 1\}^s$, do span the whole \mathbb{R}^s and the span of the $g_\sigma - g_e$, $\sigma \in \{-1, 1\}^s$, is identical to the range of Z^\top . \square

Optimality without strict complementarity

Now suppose that for a given x and corresponding Lagrange multipliers (17) holds as equality for a subset of indices $i \in \mathcal{J} \subset \{1, \dots, s\} \equiv \mathcal{I}$, which means that some Lagrange multipliers μ_j associated with the sign conditions $\bar{z}_j \geq 0$ vanish. To obtain sharper necessary as well as still sufficient second order optimality conditions one can require that the Hessian of the Lagrangian is positive semi-definite or definite when projected onto an enlarged null space of the Jacobian that is obtained by deleting the inequality constraint gradients associated with zero Lagrange multipliers [9]. In effect, this means applying the standard KKT conditions and second-order sufficient conditions (SSC) to a modified problem where the inequality constraints lacking strict complementarity are locally relaxed. For this purpose, we consider a point x such that all kinks are localized and LIKQ as well as the tangential stationarity condition are satisfied. Then, we define the set

$$\mathcal{J} \equiv \{i \in \mathcal{I} \mid \mu_i = 0\}$$

collecting the indices of all vanishing Lagrange multipliers μ_i with $\tilde{s} \equiv |\mathcal{J}|$ elements. Correspondingly, we consider the reduced vector

$$\tilde{z} \equiv (e_i^\top \bar{z})_{i \in \mathcal{J}},$$

collecting all variables $\bar{z}_i, i \in \mathcal{J}$, that are zero and have a vanishing Lagrange multiplier. Fixing $\bar{z}_i, i \in \mathcal{I} \setminus \mathcal{J}$, one can derive from the branch problems given in Eq. (15) reduced optimization problems with $(x, \bar{z}) \in \mathbb{R}^{n+\bar{s}}$ as optimization variables that are again smooth such that standard KKT conditions and SSC can be applied. For the projection of the Hessian of the Lagrangian, we derive from the matrix V as defined in Eq. (12) the correspondingly reduced version

$$\tilde{V} \equiv [V(V^\top V)^{-1}(\Sigma - L)e_i]_{i \in \mathcal{J}} \in \mathbb{R}^{n \times \bar{s}}$$

Based on these reduced quantities, we get the following results, where we assume that no Lagrange multiplier pair (λ_i, μ_i) vanishes simultaneously, which would represent an even higher degeneracy.

Proposition 1 (Dual degenerate second order optimality). *In the localized scenario assume that x satisfies LIKQ and the tangential stationarity condition such that*

$$|\lambda_i| + \mu_i > 0 \text{ for } i \in \mathcal{I}.$$

Setting $\sigma \equiv \text{sgn}(\lambda)$, we find that x can only be a local minimizer if the projected Hessian of order $n - s + \bar{s}$ given by

$$\begin{bmatrix} U^\top & 0 \\ \tilde{V}^\top & I_{\bar{s}} \end{bmatrix} \begin{bmatrix} f_{xx} + (\lambda^\top F)_{xx} & f_{x\bar{z}} + (\lambda^\top F)_{x\bar{z}} \\ f_{\bar{z}x} + (\lambda^\top F)_{\bar{z}x} & f_{\bar{z}\bar{z}} + (\lambda^\top F)_{\bar{z}\bar{z}} \end{bmatrix} \begin{bmatrix} U & \tilde{V} \\ 0 & I_{\bar{s}} \end{bmatrix} \quad (22)$$

is positive semi-definite. Moreover x must be a strict local minimizer if this projected Hessian matrix is positive definite. Here all derivatives are evaluated at $(x, 0) \in \mathbb{R}^{(n+s)}$.

Proof. Since λ_i and μ_i cannot vanish simultaneously only those branch problems defined by $\tilde{\sigma}$ are lacking strict complementarity for which $\tilde{\sigma}_j = \sigma_j \neq 0$ for some $j \in \mathcal{J}$. Setting in Eq. (16) for $j \in \mathcal{J}$ the $(s+j)$ -th row that represents an active inequality constraint with a zero Lagrange multiplier to zero corresponds to a relaxation of this constraint without changing the dimension of the Jacobian. Then, we obtain an extra null vector of the Jacobian given in Eq. (16) as

$$\begin{bmatrix} V^+(\sigma_j e_j - L e_j) \\ e_j \end{bmatrix} \in \mathbb{R}^{n+s} \quad \text{since} \quad V^\top V^+(\Sigma - L)e_j + (L - \Sigma)e_j = 0,$$

where $V^+ = V(V^\top V)^{-1}$ is the Penrose inverse of V^\top . One should note that any one of these extra nullvectors is orthogonal to the joint null vectors spanned by the columns of U padded with s trailing zeros. Moreover it is completely independent of the other components $\tilde{\sigma}_i$ with $i \neq j$ characterizing the branch problems. Hence all extended Jacobian nullspaces for the branch problems are contained in the one that is obtained by setting $\sigma_i = \text{sgn}(\lambda_i)$ for all $i \in \mathcal{I}$ as we have done in the assumptions. Its full nullspace is given by

$$\begin{bmatrix} U & \tilde{V} \\ 0 & I_{\bar{s}} \end{bmatrix} \in \mathbb{R}^{(n+\bar{s}) \times (n-s+\bar{s})}.$$

Since $\tilde{\Sigma}$ and the corresponding $\tilde{\mu}$ enter only in linear terms of the constraints of the branch problems (15) their full space Lagrangian Hessians are all identical and we obtain the maximal projected Hessian given in Eq. (22). The projected Hessians of all other branch problems are principal minors of this matrix, whose positive definiteness is therefore sufficient for optimality of the underlying non-smooth problem. \square

We have carried out this analysis in the localized scenario, but it can naturally be transformed backwards to the original setting as shown in Proposition 5.

Descent paths and the local model problem

As in smooth optimization it is natural to iteratively solve the original problem by successively computing descent steps as long as the necessary optimality conditions are violated. We demonstrate how this is possible for the fully active case $|\alpha| = s \leq n$ and leave a fully fledged algorithmic development to a subsequent paper.

Proposition 2 (Descent direction from first order violation). *Suppose the tangential stationarity or the weak form of the normal growth condition is violated at a point x where LIKQ holds and $z(x) = 0$. Then there exists a direction $\Delta x \in \mathbb{R}^n$ and a constant $c > 0$ such that*

$$\varphi(x + \Delta x \tau) = \varphi(x) - c\tau + \mathcal{O}(\tau^2).$$

Proof. If the tangential stationarity condition is violated there exists a direction Δx such that $a^\top \Delta x < 0$ and $Z\Delta x = 0$. This means for the corresponding increments $\Delta z(\tau) = z(x + \tau\Delta x)$ with a step multiplier $\tau > 0$ that by Taylor expansion of our switch equation (5)

$$\Delta z(\tau) = F(x, |\Delta z(\tau)|) = L|\Delta z(\tau)| + \mathcal{O}(\tau^2 + \|\Delta z(\tau)\|^2).$$

Since the order term is also strictly lower triangular, i.e., $\Delta z_i(\tau)$ depends only on $\Delta z_j(\tau)$ for $j < i$, one finds easily by induction that $\Delta z(\tau) = \mathcal{O}(\tau^2)$ so that again by Taylor

$$\begin{aligned} f(x + \tau\Delta x, |\Delta z(\tau)|) - f(x, 0) &= a^\top \Delta x \tau + b^\top |\Delta z(\tau)| + \mathcal{O}(\tau^2) \\ &= a^\top \Delta x \tau + \mathcal{O}(\tau^2) < 0 \quad \text{for } \tau \gtrsim 0. \end{aligned}$$

Hence one can construct a descent direction if the tangential stationarity condition is violated. Suppose we have tangential stationarity but the normal growth condition is violated even in its weak form so that for some index i

$$b_i = \frac{\partial f(x, |z|)}{\partial |z_i|} < |\lambda_i| - e_i^\top L^\top \lambda,$$

where $e_i \in \mathbb{R}^s$ is the i -th Cartesian basis vector. Set $\sigma_i \in \{-1, 1\}$ such that $\lambda_i \sigma_i \geq 0$. Then there exists a direction Δx such that $Z\Delta x = (I\sigma_i - L)e_i$. Again

we modify $x(\tau) = x + \tau\Delta x$ and obtain now for the corresponding switching equation

$$\begin{aligned}
\Delta z(\tau) &= F(x, |\Delta z(\tau)|) \\
&= \tau Z\Delta x + L|\Delta z(\tau)| + \mathcal{O}(\tau^2 + \|\Delta z(\tau)\|^2) \\
&= \tau(I\sigma_i - L)e_i + L|\Delta z(\tau)| + \mathcal{O}(\tau^2 + \|\Delta z(\tau)\|^2) \\
&= \tau(I\sigma_i - L)e_i + L|\Delta z(\tau)| + \mathcal{O}(\tau^2),
\end{aligned}$$

where the last equation follows from the fact that $\Delta z(\tau)$ is Lipschitz continuous w.r.t. τ . Now, if we substitute $\Delta z(\tau) = \sigma_i \tau e_i$ and thus $|\Delta z(\tau)| = \tau e_i$ into the above equation the error is of $\mathcal{O}(\tau^2)$ so that due to the Lipschitz continuity of the equation system and its inverse the actual solution satisfies

$$\Delta z(\tau) = \sigma_i \tau e_i + \mathcal{O}(\tau^2)$$

and consequently $|\Delta z(\tau)| = \tau e_i + \mathcal{O}(\tau^2)$. Following the usual strategy we expand the function variation along the path $x(\tau)$ as

$$\begin{aligned}
\varphi(x(\tau)) &= f(x(\tau), |\Delta z(\tau)|) \\
&= f(x(\tau), |\Delta z(\tau)|) + \lambda^\top [F(x(\tau), |\Delta z(\tau)|) - \Delta z(\tau)] \\
&= \varphi(x) + (a^\top + \lambda^\top Z)\Delta x \tau + (b^\top + \lambda^\top L)|\Delta z(\tau)| - |\lambda_i|\tau + \mathcal{O}(\tau^2) \\
&= \varphi(x) + (b_i + \lambda^\top L e_i - |\lambda_i|)\tau + \mathcal{O}(\tau^2) < 0 \quad \text{for } \tau \gtrsim 0.
\end{aligned}$$

Hence we have constructed a descent direction whenever the *normal growth* condition is not satisfied. \square

Proposition 3 (Descent parabola from second order violation). *Suppose the tangential stationarity holds but \check{H} has a negative eigenvalue at a point x where LKQ holds and $z(x) = 0$. Then there exist directions $\Delta x, \Delta'x \in \mathbb{R}^n$ and a constant $c > 0$ such that*

$$\varphi(x + \Delta x \tau + \Delta'x \tau^2) = \varphi(x) - c\tau^2 + \mathcal{O}(\tau^3).$$

Proof. Since the semi-definiteness condition is violated there exists a direction $\Delta x = Uu$ of negative curvature with some coefficient vector $u \in \mathbb{R}^{n-s}$ such that $c \equiv -u^\top \check{H}u/2 > 0$. Moreover, there exists a second vector $\Delta'x$ such that by Taylor $Z\Delta'x = -\frac{1}{2}F_{xx}(x, 0)[\Delta x, \Delta x]$. Then we find for $x(\tau) = x + \Delta x \tau + \Delta'x \tau^2$ that

$$\begin{aligned}
F(x(\tau), 0) &= F_x(x, 0)(x(\tau) - x) + \frac{1}{2}F_{xx}(x, 0)[x(\tau) - x, x(\tau) - x] + \mathcal{O}(\tau^3) \\
&= Z\Delta x \tau + Z\Delta'x \tau^2 + \frac{1}{2}F_{xx}(x, 0)[\Delta x, \Delta x] \tau^2 + \mathcal{O}(\tau^3) = \mathcal{O}(\tau^3).
\end{aligned}$$

Consequently we find for $z(\tau) = z(x(\tau))$ by Taylor expansion w.r.t. $|z|$ at 0

$$\begin{aligned}
z(\tau) &= F(x(\tau), |z(\tau)|) \\
&= F(x(\tau), 0) + L(x(\tau), 0)|z(\tau)| + \mathcal{O}(\|z(\tau)\|^2) \\
&= L(x(\tau), 0)|z(\tau)| + \mathcal{O}(\|z(\tau)\|^2) + \mathcal{O}(\tau^3).
\end{aligned}$$

Using again the strict upper triangularity of the system we find by induction that $z(\tau) = \mathcal{O}(\tau^3)$ and thus

$$\begin{aligned}
f(x(\tau), |z(\tau)|) &= f(x(\tau), 0) + \mathcal{O}(\tau^3) \\
&= f(x, 0) + a^\top \Delta x \tau + a^\top \Delta' x \tau^2 + \frac{1}{2} \Delta x^\top f_{xx}(x, 0) \Delta x \tau^2 + \mathcal{O}(\tau^3) \\
&= f(x, 0) - \lambda^\top Z \Delta x \tau - \lambda^\top Z \Delta' x \tau^2 + \frac{1}{2} \Delta x^\top f_{xx}(x, 0) \Delta x \tau^2 + \mathcal{O}(\tau^3) \\
&= f(x, 0) + \frac{1}{2} \Delta x^\top [f_{xx}(x, 0) + (\lambda^\top F(x, 0))_{xx}] \Delta x \tau^2 + \mathcal{O}(\tau^3) \\
&= f(x, 0) + \frac{1}{2} u^\top U H U u \tau^2 + \mathcal{O}(\tau^3) = f(x, 0) - c \tau^2 + \mathcal{O}(\tau^3),
\end{aligned}$$

which proves the assertion since by definition $f(x(\tau), |z(\tau)|) \equiv \varphi(x(\tau))$. \square

As in the smooth case the combined first and second order optimality conditions at some point x are equivalent to the trivial increment $\Delta x = 0$ being the minimum of a local model problem. To derive that problem we expand (5) and (6) at x, z with $z = 0$ and thus $F(x, 0) = 0$ by Taylor to $x + \Delta x$ and Δz

$$\begin{aligned}
\Delta z &\approx Z \Delta x + L(|z + \Delta z| - |z|), \\
y + \Delta y &\approx y + a^\top \Delta x + b^\top (|z + \Delta z| - |z|) + \frac{1}{2} \Delta x^\top H \Delta x.
\end{aligned}$$

Thus we obtain the local model problem of minimizing the incremental objective Δy defined by

$$\Delta z = Z \Delta x + L(|z + \Delta z| - |z|), \quad (23)$$

$$\Delta y = a^\top \Delta x + b^\top (|z + \Delta z| - |z|) + \frac{1}{2} \Delta x^\top H \Delta x. \quad (24)$$

Applying the optimality condition derived in the previous subsection the trivial increment $\Delta x = 0$ can only be a minimizer of Δy if for some $\lambda \in \mathbb{R}^s$

$$a = Z^\top \lambda, \quad b \geq |\lambda| - L^\top \lambda \quad \text{and} \quad U^\top H U \succeq 0 \quad \text{with} \quad Z U = 0.$$

Moreover, $\Delta x = 0$ must be a minimizer if the lower bounds on the components of b hold strictly and $U^\top H U$ is positive definite.

Hence the first and second order optimality condition of the original and the model problem coincide, provided H is chosen according to (14) at the current point x . The model problem can also be rewritten as

$$\min_{\Delta x} \Delta \varphi(x; \Delta x) + \frac{1}{2} \Delta x^\top H \Delta x$$

where $\Delta \varphi(x; \Delta x)$ represents the incremental piecewise linearization of φ at x as defined in [3]. In [6] an algorithm was given for the minimization of the right hand side when H is a multiple of the identity so that $\frac{1}{2} \Delta x^\top H \Delta x$ is just a proximal term in the Euclidean norm. Obviously the generalization to another ellipsoidal norm is straight forward as long as H is maintained positive definite. The efficacy of this successive piecewise linearization approach with a proximal term based on a positive definite Lagrangian Hessian approximation remains to be explored. One would hope to achieve superlinear convergence under suitable conditions on H .

4 Localization and main result

As we want to state the optimality conditions formally for the general case with $s > |\alpha|$ we need to perform the reduction to a description where all switching variables vanish at the point of potential minimality. Let $\sigma = \sigma(x)$ be fixed. Due to the continuity of our state equation $z = F(x, |z|)$ the components of the vector $\hat{z} \equiv (\sigma_i z_i)_{i \notin \alpha} \equiv (|z_i|)_{i \notin \alpha} \in \mathbb{R}^{|\sigma|}$ will keep their positive sign in some neighborhood $B \ni x$. Defining the complement as $\check{z} = (z_i)_{i \in \alpha} \in \mathbb{R}^{|\alpha|}$, we partition the argument of F and f accordingly and write $F(x, |\check{z}|, \hat{z})$ and $f(x, |\check{z}|, \hat{z})$ instead of $F(x, |z|)$ and $f(x, |z|)$. Moreover, we partition

$$\hat{F} = (\sigma_i e_i^\top F)_{i \notin \alpha} \in \mathbb{R}^{|\sigma|} \quad \text{and} \quad \check{F} = (e_i^\top F)_{i \in \alpha} \in \mathbb{R}^{|\alpha|}. \quad (25)$$

In effect we have redefined all inactive switching variables to be positive and we obtain the defining system

$$\hat{z} = \hat{F}(x, \bar{z}, \hat{z}) \quad \text{with} \quad \bar{z} = |\check{z}| \in \mathbb{R}^\alpha.$$

Being smooth and still strictly triangular with respect to the components of \hat{z} this system is uniquely solvable with respect to \hat{z} for given x and \bar{z} so that one obtains the globally unique function $\hat{z} = \hat{z}(x, \bar{z})$. By the implicit function theorem the derivatives are given by

$$\hat{z}_x \equiv \frac{\partial \hat{z}}{\partial x} = (I - \hat{F}_{\hat{z}})^{-1} \hat{F}_x \quad \text{and} \quad \hat{z}_{\bar{z}} \equiv \frac{\partial \hat{z}}{\partial \bar{z}} = (I - \hat{F}_{\hat{z}})^{-1} \hat{F}_{\bar{z}}. \quad (26)$$

Again these systems are well defined because $\hat{L} \equiv \hat{F}_{\hat{z}}$ is, like L , strictly lower triangular. Using the resulting reduced state equation one obtains for the chosen x and the correspondingly fixed $\sigma = \sigma(x)$ the smooth *localization at x* given by

$$\begin{aligned} \check{z} &= \check{F}(x, |\check{z}|, \hat{z}(x, |\check{z}|)) \\ y &= f(x, |\check{z}|, \hat{z}(x, |\check{z}|)). \end{aligned} \quad (27)$$

Clearly when the original system is piecewise linear then the reduction process and the resulting localized problem will have the same property. Moreover, then the following reduced quantities will be constant, while they are generally smooth functions of x :

$$\begin{aligned} \check{Z} &\equiv \frac{\partial}{\partial x} \check{F}(x, \bar{z}, \hat{z}(x, \bar{z})) &= \check{F}_x + \check{F}_{\bar{z}}(I - \hat{F}_{\hat{z}})^{-1} \hat{F}_x \in \mathbb{R}^{|\alpha| \times n}, \\ \check{L} &\equiv \frac{\partial}{\partial \bar{z}} \check{F}(x, \bar{z}, \hat{z}(x, \bar{z})) &= \check{F}_{\bar{z}} + \check{F}_{\hat{z}}(I - \hat{F}_{\hat{z}})^{-1} \hat{F}_{\bar{z}} \in \mathbb{R}^{|\alpha| \times |\alpha|}, \\ \check{a} &\equiv \frac{\partial}{\partial x} f(x, \bar{z}, \hat{z}(x, \bar{z})) &= f_x + f_{\hat{z}}(I - \hat{F}_{\hat{z}})^{-1} \hat{F}_x \in \mathbb{R}^n, \\ \check{b} &\equiv \frac{\partial}{\partial \bar{z}} f(x, \bar{z}, \hat{z}(x, \bar{z})) &= f_{\bar{z}} + f_{\hat{z}}(I - \hat{F}_{\hat{z}})^{-1} \hat{F}_{\bar{z}} \in \mathbb{R}^{|\alpha|}. \end{aligned} \quad (28)$$

The LIKQ condition at x requires that \check{Z} be of full row rank $|\alpha|$, which in turn is equivalent to the matrix

$$\bar{V}^\top \equiv \begin{bmatrix} \check{F}_x & \check{F}_{\hat{z}} \\ \hat{F}_x & \hat{F}_{\hat{z}} - I \end{bmatrix} \in \mathbb{R}^{s \times (n+|\sigma|)} \quad (29)$$

having full row rank s , as one can easily check by block elimination. That property yields immediately the uniqueness of the Lagrange multipliers as we will see below. Furthermore, due to the full row rank of this matrix, there exists an orthogonal matrix $\bar{U} \in \mathbb{R}^{(n+|\sigma|) \times (n+|\sigma|-s)} = \mathbb{R}^{(n+|\sigma|) \times (n-|\alpha|)}$ spanning the nullspace of \bar{V}^\top , i.e., $\bar{V}^\top \bar{U} = 0$.

The nullspace of \check{Z} is spanned by a matrix $\check{U} \in \mathbb{R}^{n \times (n-|\alpha|)}$ and the Lagrange multiplier $\check{\lambda} \in \mathbb{R}^{|\alpha|}$ of the reduced problem (27) must satisfy the tangential stationarity condition

$$f_x + f_{\hat{z}} \hat{F}_x \hat{z}_x + \check{\lambda}^\top [\check{F}_x + \check{F}_{\hat{z}} \hat{z}_x] = 0$$

with \hat{z}_x as defined in Eq. (26) and also the inequalities (18) with everything replaced by the checked quantities.

For the second order condition we differentiate once more with respect to x for fixed $\hat{z} = 0 = \bar{z}$ and $\hat{z} = \hat{z}(x, 0)$ which yields for example

$$f(x, 0, \hat{z}(x, 0))_{xx} = f_{xx} + f_{x\hat{z}} \hat{z}_x + (\hat{z}_x)^\top f_{\hat{z}x} + (\hat{z}_x)^\top f_{\hat{z}\hat{z}} \hat{z}_x$$

where the right hand side is also evaluated at $(x, 0, \hat{z}(x, 0))$. Hence we obtain the localized Hessian as

$$\check{H} = \begin{bmatrix} I & \hat{z}_x^\top \end{bmatrix} \begin{bmatrix} f_{xx} + (\check{\lambda}^\top \check{F})_{xx} & f_{x\hat{z}} + (\check{\lambda}^\top \check{F})_{x\hat{z}} \\ f_{\hat{z}x} + (\check{\lambda}^\top \check{F})_{\hat{z}x} & f_{\hat{z}\hat{z}} + (\check{\lambda}^\top \check{F})_{\hat{z}\hat{z}} \end{bmatrix} \begin{bmatrix} I \\ \hat{z}_x \end{bmatrix} \in \mathbb{R}^{n \times n} \quad (30)$$

To formulate the central result of this paper, we will also use a trunk problem similar to the one formulated already at the beginning of Sec. 3. Here, we will change the order of the variables to simplify notation later on. If x is a local minimizer of the target function $\varphi(\cdot)$ with $z_i = 0$ for all $i \in \alpha$, it must also be a local minimizer of the trunk problem

$$\begin{aligned} \min & f(x, \hat{z}(x, 0), 0) \quad \text{s.t.} \\ 0 & = \check{F}(x, \hat{z}(x, 0), 0) \quad \text{and} \quad \hat{z}(x, 0) = \check{F}(x, \hat{z}(x, 0), 0) \end{aligned}$$

which is locally smooth. For $\check{\Sigma} \in \{-1, 1\}^{|\alpha|}$, the corresponding $2^{|\alpha|}$ branch problems are given by

$$\begin{aligned} \min & f(x, \hat{z}, \bar{z}) \quad \text{s.t.} \\ \check{\Sigma} \bar{z} & = \check{F}(x, \hat{z}, \bar{z}), \quad \hat{z} = \check{F}(x, \hat{z}, \bar{z}) \quad \text{and} \quad \bar{z} \geq 0. \end{aligned} \quad (31)$$

Hence, one obtains locally smooth optimization problems in $n + s$ optimization variables. Once again, x can only be a local minimizer of the target function $\varphi(\cdot)$, if it is also a local minimizer of all branch problems. As already explained

above, the \hat{z} will stay positive in a neighborhood B of the local minimizer x such that no sign condition is required for \hat{z} .

The Jacobian of all equality and inequality constraints of Eq. (31) is given by

$$\begin{bmatrix} \check{F}_x & \check{F}_{\hat{z}} & \check{F}_{\bar{z}} - \check{\Sigma} \\ \hat{F}_x & \hat{F}_{\hat{z}} - I & \hat{F}_{\bar{z}} \\ 0 & 0 & I_{|\alpha|} \end{bmatrix} = \begin{bmatrix} \bar{V}^\top & \begin{pmatrix} \check{F}_{\bar{z}} - \check{\Sigma} \\ \hat{F}_{\bar{z}} \end{pmatrix} \\ 0 & I_{|\alpha|} \end{bmatrix} \in \mathbb{R}^{(s+|\alpha|) \times (n+s)}. \quad (32)$$

Since the matrix given in Eq. (29) has full rank s , the Jacobian of all constraints given in Eq. (32) has full rank $s + |\alpha|$ if LIKQ is fulfilled. Then, LICQ holds and standard KKT theory can be used yielding the central result of this paper, i.e., necessary and sufficient optimality conditions for a minimizer of the original nonsmooth optimization problem:

Proposition 4 (KKT + SSC in abs-normal form). *Let $\varphi : \mathbb{R}^n \mapsto \mathbb{R}$ be a nonsmooth function stated in the form (5) and (6). Suppose there are exactly $|\alpha|$ switching variables $z_i(x)$ for $i \in \alpha$ that vanish at a given point x . Moreover, assume LIKQ holds, i.e., the Jacobian matrix $\check{Z} \in \mathbb{R}^{|\alpha| \times n}$ defined in Eq. (28) has full rank $|\alpha|$. Then x can only be a local minimizer of $\varphi(\cdot)$ if there exist two unique multiplier vectors $\check{\lambda} \in \mathbb{R}^{|\alpha|}$ and $\hat{\lambda} \in \mathbb{R}^{|\sigma|}$ satisfying the KKT conditions*

Tangential stationarity:

$$[f_x, f_{\hat{z}}] = -[\check{\lambda}^\top \ \hat{\lambda}^\top] \begin{bmatrix} \check{F}_x & \check{F}_{\hat{z}} \\ \hat{F}_x & \hat{F}_{\hat{z}} - I \end{bmatrix} \in \mathbb{R}^{n+|\sigma|} \quad (33)$$

Normal growth:

$$f_{\bar{z}} \geq |\check{\lambda}^\top| - [\check{\lambda}^\top \ \hat{\lambda}^\top] \begin{bmatrix} \check{F}_{\bar{z}} \\ \hat{F}_{\bar{z}} \end{bmatrix} \in \mathbb{R}^{|\alpha|} \quad (34)$$

Positive Curvature:

$$0 \preceq \check{U}^\top \check{H} \check{U} \in \mathbb{R}^{(n-|\alpha|) \times (n-|\alpha|)} \quad (35)$$

where $0 \preceq$ denotes positive semi-definiteness of matrices.

If furthermore the $|\alpha|$ inequalities in (34) hold strictly and $\det(\check{U}^\top \check{H} \check{U}) > 0$ then x must be a local minimizer of the objective $\varphi(\cdot)$.

Proof. Since LIKQ holds, one also has that LICQ is fulfilled for the smooth branch problems. Then, the KKT theory requires that the derivative of the resulting Lagrange function with respect to x and \bar{z} must be equal to zero. This yields for the Lagrange multiplier $\check{\lambda} \in \mathbb{R}^{|\alpha|}$ and $\hat{\lambda} \in \mathbb{R}^{|\sigma|}$ of the equality and inequality constraints the equations

$$\begin{aligned} 0 &= f_x + \check{\lambda}^\top \check{F}_x + \hat{\lambda}^\top \hat{F}_x \\ 0 &= f_{\hat{z}} + \check{\lambda}^\top \check{F}_{\hat{z}} + \hat{\lambda}^\top (\hat{F}_{\hat{z}} - I), \end{aligned}$$

which are independent of the specific choice of $\check{\Sigma}$. Hence, if $\check{\lambda}$ and $\hat{\lambda}$ satisfy the tangential stationary condition (33), these necessary optimality conditions are fulfilled for all branch problems. As a further necessary first-order condition, the derivative of the corresponding Lagrange function with respect to \bar{z} must vanish. Combining this with the necessary first-order condition for the inequality constraints, one obtains for the corresponding Lagrange multiplier $\bar{\mu} \in \mathbb{R}^{|\alpha|}$ the relation

$$f_{\bar{z}} + \check{\lambda}^\top (\check{F}_{\bar{z}} - \check{\Sigma}) + \hat{\lambda}^\top \hat{F}_{\bar{z}} \equiv \bar{\mu} \geq 0. \quad (36)$$

Once more, this inequality must hold for all signature matrices and therefore also for all $\check{\Sigma}$ with

$$\check{\sigma}_i \check{\lambda}_i \geq 0 \quad \text{so that} \quad \check{\Sigma} \check{\lambda} = |\check{\lambda}|.$$

These choices impose the strongest conditions as stated also in the normal growth condition (34). Hence, if $\check{\lambda}$ fulfills this condition then also the first-order necessary condition (36).

The second order conditions follow from standard optimization theory for the smooth localization (27). \square

Provided the problems functions F and f are twice continuously differentiable it follows that the strong second order conditions are stable in that small perturbations of them in the C^2 norm still have solution points x that are strict local minimizers of the correspondingly perturbed φ .

Using the same derivation as in the subsection on the local model, one can show that localization as introduced here, i.e., the elimination of inactive switching variables, and piecewise linearization commute as illustrated in Fig. 5.

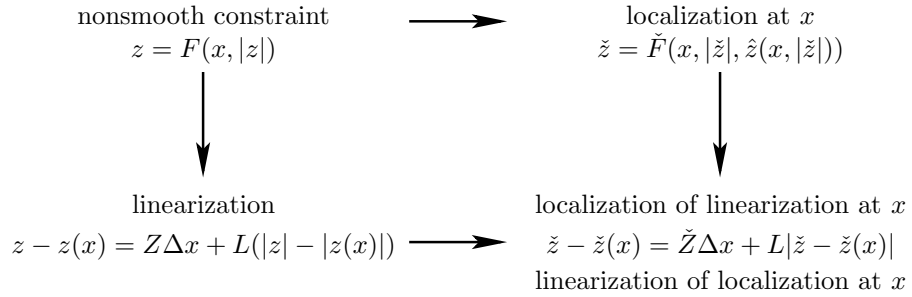


Figure 5: Commutativity of localization and linearization

Also for the general case considered in this section, it is possible to formulate second-order sufficient conditions without the assumption of strict complementarity. For this purpose assume that for a given x and corresponding Lagrange multipliers the normal growth condition (34) holds as equality for a subset of indices $\tilde{\alpha} \subset \alpha$ with $|\tilde{\alpha}|$ elements. It follows that $\bar{z}_i = 0 = \bar{\mu}_i$, where $\bar{\mu}_i$ denotes the

corresponding Lagrange multiplier of the i -th inequality for all $i \in \tilde{\alpha}$. Similar to the approach in Sec. 3, we fix all $\tilde{z}_i, i \in \alpha \setminus \tilde{\alpha}$, to obtain for

$$\tilde{\tilde{z}} \equiv (e_i^\top \tilde{z})_{i \in \tilde{\alpha}}$$

further reduced smooth branch problems with only $(x, \hat{z}, \tilde{\tilde{z}}) \in \mathbb{R}^{n+|\sigma|+|\tilde{\alpha}|}$ as optimization variables and

$$\tilde{H} = \begin{bmatrix} (f + \check{\lambda}^\top \check{F} + \hat{\lambda}^\top \hat{F})_{xx} & (f + \check{\lambda}^\top \check{F} + \hat{\lambda}^\top \hat{F})_{x\hat{z}} & (f + \check{\lambda}^\top \check{F} + \hat{\lambda}^\top \hat{F})_{x\tilde{\tilde{z}}} \\ (f + \check{\lambda}^\top \check{F} + \hat{\lambda}^\top \hat{F})_{\hat{z}x} & (f + \check{\lambda}^\top \check{F} + \hat{\lambda}^\top \hat{F})_{\hat{z}\hat{z}} & (f + \check{\lambda}^\top \check{F} + \hat{\lambda}^\top \hat{F})_{\hat{z}\tilde{\tilde{z}}} \\ (f + \check{\lambda}^\top \check{F} + \hat{\lambda}^\top \hat{F})_{\tilde{\tilde{z}}x} & (f + \check{\lambda}^\top \check{F} + \hat{\lambda}^\top \hat{F})_{\tilde{\tilde{z}}\hat{z}} & (f + \check{\lambda}^\top \check{F} + \hat{\lambda}^\top \hat{F})_{\tilde{\tilde{z}}\tilde{\tilde{z}}} \end{bmatrix} \in \mathbb{R}^{q \times q} \quad (37)$$

with $q \equiv n + |\sigma| + |\tilde{\alpha}|$ as Hessian of the corresponding Lagrangian when differentiated with respect to the optimization variables. Defining

$$\tilde{\tilde{V}} = \left[\bar{V}(\bar{V}^\top \bar{V})^{-1} \begin{pmatrix} \check{\Sigma} - \check{F}_{\tilde{z}} \\ -\hat{F}_{\tilde{z}} \end{pmatrix} e_i \right]_{i \in \tilde{\alpha}} \in \mathbb{R}^{(n+|\sigma|) \times |\tilde{\alpha}|},$$

we get the following results.

Proposition 5 (Dual degenerate second-order optimality for the general case). *Assume that exactly $|\alpha|$ switching variables $z_i(x)$ for $i \in \alpha$ vanish at a given point x . Furthermore suppose that x satisfies LIKQ and the KKT conditions (33)–(34) hold such that*

$$|\check{\lambda}_i| + \bar{\mu}_i > 0 \text{ for } i \in \alpha.$$

Setting $\check{\sigma} \equiv \mathbf{sgn}(\check{\lambda})$, we find that x can only be a local minimizer if the projected Hessian of order $n - |\alpha| + |\tilde{\alpha}|$ given by

$$\begin{bmatrix} \bar{U}^\top & 0 \\ \tilde{\tilde{V}}^\top & I_{|\tilde{\alpha}|} \end{bmatrix} \tilde{H} \begin{bmatrix} \bar{U} & \tilde{\tilde{V}} \\ 0 & I_{|\tilde{\alpha}|} \end{bmatrix} \quad (38)$$

is positive semi-definite, where \tilde{H} denotes the Hessian of the Lagrangian as stated in Eq. (37) of the doubly reduced branch problem with the optimization variables $(x, \hat{z}, \tilde{\tilde{z}}) \in \mathbb{R}^{n+|\sigma|+|\tilde{\alpha}|}$. Moreover, x must be a strict local minimizer if this projected Hessian matrix is positive definite. Here all derivatives are evaluated at $(x, \hat{z}(x, 0), 0) \in \mathbb{R}^{n+|\sigma|+|\tilde{\alpha}|}$.

Proof. By assumption no Lagrange multiplier pair $(\check{\lambda}_i, \bar{\mu}_i)$ vanishes simultaneously, which would represent an even higher degeneracy. Then only those branch problems defined by $\check{\sigma}$ are lacking strict complementarity for which $\check{\sigma}_j = \check{\sigma}_j \neq 0$ for some $j \in \tilde{\alpha}$. Setting in Eq. (32) for $j \in \tilde{\alpha}$ the $(s+j)$ -th row that represents an active inequality constraint with a zero Lagrange multiplier to zero corresponds to a relaxation of this constraint without changing the dimension of the Jacobian. Then, we obtain an extra null vector of the Jacobian given in Eq. (32) as

$$\begin{bmatrix} \bar{V}^+ \begin{pmatrix} \check{\Sigma} - \check{F}_{\tilde{z}} \\ -\hat{F}_{\tilde{z}} \end{pmatrix} e_j \\ e_j \end{bmatrix} \in \mathbb{R}^{n+s} \text{ since } \bar{V}^\top \bar{V}^+ \begin{pmatrix} \check{\Sigma} - \check{F}_{\tilde{z}} \\ -\hat{F}_{\tilde{z}} \end{pmatrix} e_j + \begin{pmatrix} \check{F}_{\tilde{z}} - \check{\Sigma} \\ \hat{F}_{\tilde{z}} \end{pmatrix} e_j = 0,$$

where $\bar{V}^+ = \bar{V}(\bar{V}^\top \bar{V})^{-1}$ is the Penrose inverse of \bar{V}^\top . Similar to the fully localized scenario one can now argue that the full nullspace is given by

$$\begin{bmatrix} \bar{U} & \bar{V} \\ 0 & I_{|\bar{\alpha}|} \end{bmatrix} \in \mathbb{R}^{(n+|\sigma|+|\bar{\alpha}|) \times (n-|\alpha|+|\bar{\alpha}|)}.$$

Since $\check{\Sigma}$ and the corresponding $\bar{\mu}$ enter only in linear terms of the constraints of the branch problems (31) their full space Lagrangian Hessians are all identical and we obtain the maximal projected Hessian given in Eq. (37). The projected Hessians of all other branch problems are principal minors of this matrix, whose positive definiteness is therefore sufficient for optimality of the underlying non-smooth problem. \square

Application to Nesterov variants

We conclude the paper by illustrating the applicability of our optimality conditions on the first and second Nesterov examples. On the first example (1) one sees from (9) that $a(x)^\top$ can only be a linear combination of some or all rows of $Z = F_x$ as required by tangential stationarity if $a(x) = 0$ which happens exactly when $x_1 = 1$. Then still by tangential stationarity we must also have $f_{\bar{z}} = 0$ but that is impossible unless this part of $b = (1 \cdots 1)$ is empty as $\alpha = \{1, 2, \dots, n-1\}$. In other words all kinks must be active and then the normal growth condition is strongly satisfied with $f_{\bar{z}} = b > 0$.

Then the nullspace of the active Jacobian is spanned by the first Cartesian basis vector e_1 and since the Lagrange multipliers are all zero we get the projected Hessian

$$e_1^\top f_{xx} e_1 = \frac{\partial^2 f(x, 0)}{\partial x_1^2} = \frac{\partial^2}{\partial x_1^2} \left[\frac{1}{4}(x_1 - 1)^2 \right] = \frac{1}{2} > 0.$$

Hence the only point satisfying the tangential stationarity and normal growth condition satisfies the latter strongly and has a positive definite projected Hessian. Hence, it is a strict minimizer.

Finally, we look at the second Nesterov example φ_2 . As observed in [7] it has apart from the global minimizer $x = (1 \cdots 1)$ exactly $2^{n-1} - 1$ Clarke stationary points which are characterized by exactly one x_i with $i < n$ being zero and all linking switches σ_{n+j} for $j = 1 \dots n-1$ being active. All other points in \mathbb{R}^n are not Clarke stationary and must therefore fail the new first order necessary conditions. We now show this directly, although unfortunately the argument is not as straightforward as one might hope. At any one of the Clarke stationary points the active set is of the form $\alpha = \{i, n+1, \dots, 2n-1\}$ so that $|\alpha| = n$. Hence the matrix (29) is square of order s and by LIKQ also nonsingular. Therefore the tangential stationarity condition must be trivially satisfied. This corresponds to the observation that the reduced Jacobian is square and nonsingular, which we made already in the first section when establishing LIKQ for the second example. However, the normal growth condition is violated as shown below.

Lemma 5 (Detection of nonoptimality on second Nesterov variant). *At a stationary point of (2) with $x_i = 0, i \in \{1, \dots, n-1\}$, the corresponding component of $f_{\bar{z}}$ is zero but the corresponding lower bound in the normal growth condition is positive.*

Proof. Because the inactive switches given by the indices $\{1, \dots, i-1, i+1, \dots, n\}$ do not depend on any other switches we have $\hat{F}_{\bar{z}} = 0$. Moreover, because only the i -th switch enters into an active switch, namely the $(i+1)$ -th, we have $\tilde{F}_{\bar{z}} = -2e_{i+1}e_1^\top$. Finally, we note that $f_{\bar{z}} = (0, 1, \dots, 1)$ and obtain for the normal growth condition

$$(0, 1 \dots 1) \geq |\check{\lambda}^\top| + 2\check{\lambda}_{i+1}e_1^\top \Rightarrow 0 \geq |\check{\lambda}_1| + 2\check{\lambda}_{i+1}.$$

Hence we may invalidate the required normal growth by showing that the Lagrange multiplier $\check{\lambda}_1$ is greater than zero and $\check{\lambda}_{i+1} = 0$. Since the lower right block in (29) is the identity one easily can block eliminate $\hat{\lambda}$ and obtains for $\check{\lambda}$ from the Tangential Stationarity (33) the system of equations

$$-\check{\lambda}^\top \begin{bmatrix} \tilde{F}_x + \tilde{F}_{\bar{z}}\hat{F}_x \end{bmatrix} = f_x + f_{\bar{z}}\hat{F}_x = \frac{1}{4}e_1^\top \quad (39)$$

which corresponds exactly to the reduction discussed above. For the second Nesterov variant, one obtains

$$\tilde{F}_x = \begin{bmatrix} 0 & e_{i-1}^\top \\ 0 & I_{n-1} \end{bmatrix} \quad \text{and} \quad \tilde{F}_{\bar{z}}\hat{F}_x = -2 \begin{pmatrix} 0 & 0 & 0 & 0 \\ I_{i-2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & I_{n-i-1} & 0 \end{pmatrix}.$$

Hence the set of equations is given by

$$\begin{aligned} \check{\lambda}_i &= \frac{1}{4}, & \check{\lambda}_{i+1} &= 0, \\ -2\check{\lambda}_j + \check{\lambda}_{j+1} &= 0 & j &= 1, \dots, i-1, \\ -2\check{\lambda}_j + \check{\lambda}_{j+1} &= 0 & j &= i+1, \dots, n-1. \end{aligned}$$

Using an induction, it follows from $\check{\lambda}_{i+1} = 0$ that $\check{\lambda}_1 = 2^{-(i+1)}$. this yields

$$0 \geq |\check{\lambda}_1| + 2\check{\lambda}_{i+1} = 2^{-(i+1)} > 0$$

and therefore a contradiction, that shows that the normal growth condition is violated. \square

5 Summary and Outlook

We have considered local optimality conditions for piecewise smooth objective functions that can be conveniently represented in the abs-normal form (5), (6). This structure is obvious not only for Nesterov's Rosenbrock variants (0), (1),

(2) but also for most of the nonsmooth test problems in the literature. Generally, a local piecewise linearization represented by the two matrices Z, L and three vectors a, b and c can be generated by an extension of automatic or algorithmic differentiation (AD) (see, e.g., [17]). Generically they satisfy at all points the Linear Independence Kink Qualification (LIKQ) and may be assumed nonredundant without loss of generality. As a side benefit we can explicitly calculate the classical $\mathcal{V}\mathcal{U}$ decomposition into the nonsmooth normal space \mathcal{V} and the smooth tangential space \mathcal{U} . More importantly, these data turn out to be sufficient to verify necessary first order optimality conditions, which consist of equalities representing tangential stationarity and inequalities representing normal growth. It is possibly a little surprising that checking these conditions requires only straightforward linear algebra and has no combinatorial aspect whatsoever.

On piecewise linear problems like the second Nesterov variant or when there are n active kinks, these necessary conditions are also sufficient. Otherwise the Lagrangian multipliers and the nullspace of the tangential stationarity condition allow again by AD the computation of a projected Hessian, whose positive semidefiniteness is necessary and whose positive definiteness in combination with strict normal growth is sufficient for strong local optimality. All these theoretical results are verified on the three Nesterov variants, where LIKQ holds globally and there are no gaps between necessary and sufficient optimality since the normal growth is always strict and projected Hessians are always positive definite. The classical Clarke and Mordukhovich stationarity conditions are stronger than the requirement of tangential stationarity but still weaker than its combination with normal growth.

Just as in the classical KKT theory one can formulate necessary and sufficient optimality conditions without strict complementarity (here strict normal growth) by requiring positive definiteness of the Lagrangian Hessian on the larger tangent space restricted only by equalities and the active inequality constraints whose Lagrange multipliers are nonzero. Obviously several other degeneracies may arise that one might be able to handle with limited additional complexity. After all, linear optimization is a rather special case of the theory developed here.

When either of the three parts of the necessary optimality condition is violated one can construct a descent direction or parabolic arc as demonstrated in Propositions 2 and 3. This observation suggests the design of an iterative optimizer based on successive piecewise linearizations with a proximal term defined by a positive definite approximation to the Hessian of the Lagrangian. Under suitable assumptions one should be able to establish local and superlinear convergence, which would definitely represent an algorithmic advance for nonsmooth optimization methods. Naturally such a method would use much more information than is available to the current front runners such as the bundle methods under the classical oracle paradigm [13]. However, we have argued in [6] that this paradigm is not very realistic and that usually either a lot more – or possibly a lot less – information about the objective function is available.

There are many other possible extensions of the theory presented here. One is the explicit treatment of piecewise smooth constraints, which can be theoretically incorporated by l_∞ or ℓ_1 penalization into the unconstrained framework. However, the penalty multiplier may be hard to determine and a lot of structure is likely to be lost. We have little doubt that Proposition 4 can be extended to the constrained case, of course involving another set of (more conventional) Lagrange multipliers. One would also wish to cover the optimization of the maxima of implicit functions, which includes the spectral radii of symmetric matrices, a problem that has received a lot of attention in the nonsmooth optimization literature. The extension to Lipschitzian but not piecewise smooth functions like the Euclidean norm is a real challenge, which may or may not be worthwhile attacking. Certainly superlinearly convergent algorithms seem to be out of the question in this much more general scenario.

Still within the framework considered here one may ask the questions, whether and where such functions φ are locally convex. Smooth optimality conditions for local minima usually combine a stationarity condition with a convexity condition. However, we noted that even piecewise linear objectives with strict local minima need not be convex in their neighborhood as it is the case for the gradient cube example if $0 < \varepsilon < 1$ and $n = 2$. Nevertheless, in some applications like multiphase equilibria of mixed fluids, lack of convexity may lead to the instability of single phase equilibria. Therefore we will examine conditions for convexity in the vicinity of a given point, irrespective of whether the point is stationary or not in a forthcoming paper titled *First and second order conditions for piecewise smooth objective functions*. Like for optimality, one can obtain necessary first order conditions and necessary or sufficient second order conditions based on the additional information that is available through the abs-normal form. However, our preliminary investigations suggest that, in sharp contrast to the polynomial nature of the SSC+KKT optimality conditions, even the test for first order convexity, i.e. convexity of the local piecewise approximation might be NP hard. Moreover, we conjecture that first order convexity is equivalent to subgradient regularity in the sense of Clarke, which is a requirement for the partial smoothness concept of Lewis.

Acknowledgment

The authors greatly benefited from many discussions with Jens-Uwe Bernt in the context of his Diploma project. The proof of Lemma 4 was partly suggested by Siegfried Rump and benefited from communications with Manuel Radons and Hermann Mena. They are also thankful to the constructive reviews and suggestions of the three referees, which greatly improved the final version.

References

- [1] R.H. Bartels and A.R. Conn. Linearly constrained discrete l_1 problems. *ACM Transactions on Mathematical Software*, 6:594–608, 1980.
- [2] R.H. Bartels, A.R. Conn, and J.W. Sinclair. Minimization techniques for piecewise differentiable functions: the ℓ_1 solution to an overdetermined linear system. *SIAM Journal on Numerical Analysis*, 15:224–241, 1978.
- [3] A. Griewank. On stable piecewise linearization and generalized algorithmic differentiation. *Opt. Meth. and Softw.*, 28(6):1139–1178, 2013.
- [4] A. Griewank, J.-U. Bernt, M. Randons, and T. Streubel. Solving piecewise linear equations in abs-normal form. *Lin. Algebra and its Appl.*, 471:500–530, 2015.
- [5] A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM, 2008.
- [6] A. Griewank, A. Walther, S. Fiege, and T. Bosse. On Lipschitz optimization based on gray-box piecewise linearization. *Mathematical Programming Series A*, 2015. To appear, available online, 10.1007/s10107-015-0934-x.
- [7] M. Gürbüzbalaban and M.L. Overton. On Nesterov’s nonsmooth Chebyshev-Rosenbrock functions. *Nonlinear Anal: Theory, Methods & Appl.*, 75(3):1282–1289, 2012.
- [8] X. Huang, J. Xu, and S. Wang. Exact penalty and optimality condition for nonseparable continuous piecewise linear programming. *Journal of Optimization Theory and Applications*, 155(1):145–164, 2012.
- [9] F. Jarre and J. Stoer. *Optimierung*. Springer, 2004.
- [10] A.S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2002.
- [11] R. Mifflin and C. Sagastizábal. On \mathcal{VU} -theory for functions with primal-dual gradient structure. *SIAM Journal on Optimization*, 11(2):547–571, 2000.
- [12] R. Mifflin and C. Sagastizábal. Primal-dual gradient structured functions: second-order results; links to epi-derivatives and partly smooth functions. *SIAM Journal on Optimization*, 13(4):1174–1194, 2003.
- [13] R. Mifflin and C. Sagastizábal. A science fiction story in nonsmooth optimization originating at IIASA. *Documenta Mathematica*, Extra Vol.:291–300, 2012.
- [14] B. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*. Springer, 2006.

- [15] Y. Nesterov. Lexicographic differentiation of nonsmooth functions. *Math. Program.*, 104(2-3):669–700, 2005.
- [16] S. Scholtes. *Introduction to Piecewise Differentiable Functions*. Springer, 2012.
- [17] A. Walther and A. Griewank. *Combinatorial Scientific Computing*, chapter Getting Started with ADOL-C, pages 181–202. Chapman-Hall CRC Computational Science, 2012.