SM Electronic supplementary material

Supplementary material to the article "On the structure of regularization paths for piecewise differentiable regularization terms" by B. Gebken, K. Bieker and S. Peitz (2022).

SM.1 Piecewise differentiable functions

In the following, we will define piecewise differentiability and state the main results that we use throughout this article. For a more detailed introduction into the topic, we refer to [1]. Let $U \subseteq \mathbb{R}^n$ be open.

Definition SM1 Let $g : U \to \mathbb{R}$ be continuous and $g_i : U \to \mathbb{R}$, $i \in \{1, \ldots, k\}$, be a set of r-times continuously differentiable (or C^r) functions for $r \in \mathbb{N} \cup \{\infty\}$. If $g(x) \in \{g_1(x), \ldots, g_k(x)\}$ for all $x \in U$, then g is *piecewise r-times differentiable* (or a *PC^r-function*). In this case, $\{g_1, \ldots, g_k\}$ is called a *set of selection functions* of g.

When working with PC^r-functions in a local sense, it is useful to only consider the selection functions that have an impact on the local behavior around a given point.

Definition SM2 Let $g : U \to \mathbb{R}$ be a PC^r-function and let $\{g_1, \ldots, g_k\}$ be a set of selection functions of g. Then

$$I(x) := \{i \in \{1, \ldots, k\} : g(x) = g_i(x)\}$$

is the *active set* at $x \in U$. A selection function g_i is called *active at x* if $i \in I(x)$.

From the continuity of selection functions it follows that for any $x^0 \in \mathbb{R}^n$, there is an open neighborhood $U' \subseteq U$ of x^0 such that

$$g(x) \in \{g_i(x) : i \in I(x^0)\} \quad \forall x \in U'. \tag{SM1}$$

But note that not all active selection functions are necessarily required for the local representation of g around x^0. For example, if a selection function is only active in x^0 and nowhere else, then it can be neglected. Hence, there is also the following, stricter definition of activity. To this end, for a set $A \subseteq \mathbb{R}^n$, we denote by $\mathrm{cl}(A)$ the *closure* and by $\mathrm{int}(A)$ the *interior* of A with respect to the natural topology on $\mathbb{R}^n$.

Definition SM3 Let $g : U \to \mathbb{R}$ be a PC^r-function and let $\{g_1, \ldots, g_k\}$ be a set of selection functions of g. Then

$$I^e(x) := \{i \in \{1, \ldots, k\} : x \in \mathrm{cl}(\mathrm{int}(\{y \in U : g(y) = g_i(y)\}))\}$$

is the *essentially active set* at $x \in U$. A selection function g_i is called *essentially active at x* if $i \in I^e(x)$.

Due to continuity we have $I^e(x) \subseteq I(x)$ for all $x \in \mathbb{R}^n$. By definition, if a selection function g_i is essentially active at some point, then $\mathrm{int}(\{y \in U : g(y) = g_i(y)\})$ is non-empty. In other words, there is an open subset of U on which g behaves like g_i. The following lemma shows that locally, a given set of selection functions can always be reduced to those that are essentially active.

Lemma SM1 Let $g : U \to \mathbb{R}$ be a PC^r-function and let $\{g_1, \ldots, g_k\}$ be a set of selection functions of g. Then for any $x^0 \in U$, there is an open neighborhood $U' \subseteq U$ of x^0 such that $\{g_i : i \in I^e(x^0)\}$ is a set of selection functions of the restriction $g|_{U'}$ of g to U'.

Proof Proposition 2.22 in [2]. □

Although we only assumed continuity in the definition of PC^r-functions, it is possible to show the following, stronger result. To this end, let $\|\cdot\|$ be any norm on $\mathbb{R}^n$.

Lemma SM2 Let $g : U \to \mathbb{R}$ be a PC^r-function. Then g is locally Lipschitz continuous, i.e., for every $x \in U$, there is an open neighborhood $U' \subseteq U$ of x and some $L > 0$ such that

$$|g(y) - g(z)| \leq L\|y - z\| \quad \forall y, z \in U'.$$

Proof Corollary 4.1.1 in [1]. □

While PC^r-functions are generally nonsmooth, the previous lemma allows us to use the so-called *Clarke subdifferential* from nonsmooth analysis to obtain first-order approximations. To this end, for $A \subseteq \mathbb{R}^n$ let $\mathrm{conv}(A)$ be the *convex hull of* A. For a general locally Lipschitz continuous function $g : U \to \mathbb{R}$, let $\Omega \subseteq U$ be the set of points in which g is not differentiable. Then the (Clarke) subdifferential of g at $x \in U$ can be defined as

$$\partial g(x) := \mathrm{conv}\left(\left\{\xi \in \mathbb{R}^n : \exists (x^j)_{j\in\mathbb{N}} \in \mathbb{R}^n \setminus \Omega \text{ with } \lim_{j\to\infty} x^j = x \text{ and } \lim_{j\to\infty} \nabla g(x^j) = \xi\right\}\right). \tag{SM2}$$

If g is continuously differentiable in x, then $\partial g(x) = \{\nabla g(x)\}$. Although the subdifferential is a set, it behaves similarly to the standard derivative. For example, there are generalized versions of the chain rule and the mean-value theorem. Furthermore, as we will see later, it can be used to obtain optimality conditions. For a more detailed introduction into nonsmooth analysis, we refer to [3, 4].

In practice, computing the Clarke subdifferential of an arbitrary locally Lipschitz function can be difficult (cf. Section 3.3 in [5]). But fortunately, for the special case of PC^r-functions, there is a simpler expression for the Clarke subdifferential in terms of the gradients of the selection functions. More precisely, we can use the following result.

Lemma SM3 Let $g : U \to \mathbb{R}$ be a PC^r-function and let $\{g_1, \ldots, g_k\}$ be a set of selection functions of g. Then

$$\partial g(x) = \mathrm{conv}(\{\nabla g_i(x) : i \in I^e(x)\}) \quad \forall x \in U.$$

Proof Proposition 4.3.1 in [1]. □

By the previous result, knowing the classical gradients of all essentially active selection functions is sufficient to obtain the exact Clarke subdifferential. In particular, since the number of selection functions is finite by definition, it follows that subdifferentials of PC^r-functions are always convex polytopes.

We conclude the introduction to PC^r-functions with some simple examples.

*Example 1*a) Consider the ℓ^1-norm on $\mathbb{R}^n$, i.e.,

$$g : \mathbb{R}^n \to \mathbb{R}, \; x \mapsto \|x\|_1 := |x_1| + \cdots + |x_n|.$$

Then

$$g(x) \in \left\{ \sum_{i=1}^{n} \sigma_i x_i : \sigma \in \{-1, 1\}^n \right\},$$

so g is PC^∞ with selection functions

$$g_\sigma : \mathbb{R}^n \to \mathbb{R}, \quad x \mapsto \sum_{i=1}^{n} \sigma_i x_i \quad \text{for} \quad \sigma \in \{-1, 1\}^n.$$

For $x^0 \in \mathbb{R}^n$, the corresponding set of essentially active selection functions in x^0 is given by

$$\left\{ g_\sigma : \sigma_i \begin{cases} = \mathrm{sign}(x_i^0), & \text{if } x_i^0 \neq 0 \\ \in \{-1, 1\}, & \text{if } x_i^0 = 0 \end{cases}, i \in \{1, \ldots, n\} \right\}.$$

Therefore, the Clarke subdifferential of g at $x^0 \in \mathbb{R}^n$ is given by

$$\partial g(x^0) = \left\{ \xi \in \mathbb{R}^n : \xi_i \begin{cases} = \mathrm{sign}(x_i^0), & \text{if } x_i^0 \neq 0 \\ \in [-1, 1], & \text{if } x_i^0 = 0 \end{cases}, i \in \{1, \ldots, n\} \right\}.$$

b) As an example for a function that is differentiable almost everywhere but not PC^r (for $r > 1$), consider

$$g : \mathbb{R} \to \mathbb{R}, \; x \mapsto \sqrt{|x|}.$$

Although g is continuous everywhere and differentiable outside of 0, it is not PC^r (since it is not locally Lipschitz continuous in 0).

SM.2 Multiobjective optimization

In the following, we will give a brief introduction to (nonsmooth) multiobjective optimization. Due to the context of our paper, we only consider problems with two objectives here. For a more detailed introduction in the general case, we refer to [6–8].

Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$. The task of simultaneously minimizing f and g is denoted as

$$\min_{x \in \mathbb{R}^n} \begin{pmatrix} f(x) \\ g(x) \end{pmatrix}$$

and is called a *multiobjective optimization problem* (MOP). The goal of multiobjective optimization is to find the so-called *Pareto set*, which is defined as follows:

Definition SM4 A point $x \in \mathbb{R}^n$ is called *Pareto optimal* if there is no $y \in \mathbb{R}^n$ with

$$f(y) < f(x) \text{ and } g(y) \leq g(x) \qquad \text{or} \qquad f(y) \leq f(x) \text{ and } g(y) < g(x).$$

The set of all Pareto optimal points is the *Pareto set*. Its image under the objective vector (f, g), i.e., the set $\{(f(x), g(x))^\top : x \text{ is Pareto optimal}\} \subseteq \mathbb{R}^2$, is the *Pareto front*.

There are various different methods for solving nonsmooth MOPs, see, e.g., [9–11]. If both f and g are locally Lipschitz continuous, then the following theorem yields a necessary condition for Pareto optimality based on the Clarke subdifferentials (cf. (SM2)).

Theorem SM1 *Let $x \in \mathbb{R}^n$ be Pareto optimal. Then*

$$0 \in \operatorname{conv}(\partial f(x) \cup \partial g(x)). \tag{SM3}$$

Proof Theorem 12 in [8]. □

Since (SM3) is only a necessary condition, we make the following definition:

Definition SM5 A point $x \in \mathbb{R}^n$ is called *Pareto critical* if it satisfies (SM3). The set of all Pareto critical points is the *Pareto critical set*, denoted by P_c.

The Pareto critical set is a superset of the actual Pareto set. If f and g are convex, then (SM3) is also sufficient, so the Pareto critical set coincides with the set of Pareto optimal points in that case (cf. Theorem 3.2.11 in [6]).

Using a result about the convex hull of the union of convex sets (cf. Lemma 5.29 in [12]), (SM3) is equivalent to

$$\exists \alpha_1, \alpha_2 \geq 0, \xi^1 \in \partial f(x), \xi^2 \in \partial g(x) : \alpha_1 \xi^1 + \alpha_2 \xi^2 = 0, \alpha_1 + \alpha_2 = 1. \quad \text{(SM4)}$$

In accordance with the smooth case, we will refer to such α_1 and α_2 as *KKT multipliers of f and g in x*, respectively. Note that (SM4) implies

$$0 \in \{\lambda_1 \xi^1 + \lambda_2 \xi^2 : \lambda_1, \lambda_2 \in \mathbb{R}, \lambda_1 + \lambda_2 = 1\},$$

where the right-hand side is a so-called *affine space*. The properties of such spaces are analyzed in the area of *affine geometry*.

SM.3 Affine geometry

In the following, we will introduce the basic concepts of affine geometry and affine spaces which we will use in this article. For further details on this topic, we refer to [13–15].

Definition SM6 a) Let $k \in \mathbb{N}$ and $a^i \in \mathbb{R}^n$, $i \in \{1, \ldots, k\}$. Let $\lambda \in \mathbb{R}^k$ with $\sum_{i=1}^k \lambda_i = 1$. Then $\sum_{i=1}^k \lambda_i a^i$ is an *affine combination of* $\{a^1, \ldots, a^k\}$.

b) Let $E \subseteq \mathbb{R}^n$. Then $\operatorname{aff}(E)$ is the set of all affine combinations of elements of E, called the *affine hull of E*. Formally,

$$\operatorname{aff}(E) := \left\{ \sum_{i=1}^k \lambda_i a^i : k \in \mathbb{N}, a^i \in E, \lambda_i \in \mathbb{R}, i \in \{1, \ldots, k\}, \sum_{i=1}^k \lambda_i = 1 \right\}.$$

c) Let $E \subseteq \mathbb{R}^n$. If $\operatorname{aff}(E) = E$, then E is called an *affine space*.

Affine spaces can be thought of as linear spaces that were translated away from the origin. More precisely, if E is an affine space and $a' \in E$, then the set

$$E - a' = \{a - a' : a \in E\} =: V \quad \text{(SM5)}$$

is a linear subspace of $\mathbb{R}^n$. (Note that V does not depend on the choice of a'). This allows for the definition of *affine independence*.

Definition SM7 Let E be an affine space and let $a^i \in E$, $i \in \{1, \dots, k\}$. Then the set $\{a^1, \dots, a^k\}$ is called *affinely independent* if $\{a^i - a^j : i \in \{1, \dots, k\} \setminus \{j\}\}$ is linearly independent for some $j \in \{1, \dots, k\}$.

If the condition in the previous definition holds for some $j \in \{1, \dots, k\}$, then it automatically holds for all $j \in \{1, \dots, k\}$ (cf. Lemma 2.4 in [14]). As for linear independence, affine independence is related to uniqueness of the coefficients of affine combinations.

Lemma SM4 Let E be an affine space and let $a^i \in E$, $i \in \{1, \dots, k\}$. Let $a \in \mathrm{aff}(\{a^1, \dots, a^k\})$ and $\lambda \in \mathbb{R}^k$ such that $a = \sum_{i=1}^k \lambda_i a^i$ and $\sum_{i=1}^k \lambda_i = 1$. Then $\{a^1, \dots, a^k\}$ is affinely independent if and only if λ is unique.

Proof Lemma 2.5 in [14]. □

Furthermore, it is possible to assign a dimension to an affine space and define *affine bases* (also known as *affine frames*).

Definition SM8 Let $k \in \mathbb{N}$ and let E be an affine space with corresponding linear subspace V (cf. (SM5)).

a) Let $a^i \in E$, $i \in \{1, \dots, k\}$. Then $\{a^1, \dots, a^k\}$ is called an *affine basis of* E if $\{a^i - a^j : i \in \{1, \dots, k\} \setminus \{j\}\}$ is a basis of V for some $j \in \{1, \dots, k\}$.
b) The *(affine) dimension* of E is the dimension of V, denoted by $\mathrm{affdim}(E)$.

The previous definition implies that an affine basis consists of $\mathrm{affdim}(E)+1$ elements. It is easy to show that if $\{a^1, \dots, a^k\}$ forms an affine basis of E, then $\mathrm{aff}(\{a^1, \dots, a^k\}) = E$ and combined with Lemma SM4, we obtain that every element in E has a unique representation as an affine combination of $\{a^1, \dots, a^k\}$.

The main reason why we need affine geometry in this article is *Carathéodory's theorem*, which gives us an upper bound for the number of elements we have to consider when computing convex hulls:

Theorem SM2 *Let A be a finite subset of $\mathbb{R}^n$. Then every element in* $\mathrm{conv}(A)$ *can be written as a convex combination of* $\mathrm{affdim}(\mathrm{aff}(A)) + 1$ *elements of A.*

Proof Theorem 3.1 in [14]. □

Carathéodory's theorem can be used to lower the number of selection functions we have to consider for the computation of the Clarke subdifferential $\partial g(x)$ of g.

Finally, the concept of affine spaces enables us to introduce a useful definition of the interior and the boundary of “low-dimensional” subsets of $\mathbb{R}^n$.

Definition SM9 Let $A \subseteq \mathbb{R}^n$ and let $\operatorname{aff}(A)$ be endowed with the subspace topology of $\mathbb{R}^n$. Then the *relative interior of* A, denoted by $\operatorname{ri}(A)$, is the interior of A in $\operatorname{aff}(A)$, i.e.,

$$\operatorname{ri}(A) := \{x \in A : \exists U \subseteq \mathbb{R}^n \text{ open with } x \in U \text{ and } U \cap \operatorname{aff}(A) \subseteq A\}.$$

The *relative boundary of* A is the set $\operatorname{rb}(A) := \operatorname{cl}(A) \setminus \operatorname{ri}(A)$, where $\operatorname{cl}(A)$ is the closure of A in $\mathbb{R}^n$.

In the case of convex polytopes, i.e., convex hulls of a finite number of points, the relative interior and boundary can be expressed in terms of the coefficients of convex combinations:

Lemma SM5 Let $A = \operatorname{conv}(\{a^1, \dots, a^k\}) \subseteq \mathbb{R}^n$ with $k \in \mathbb{N}$. Then

$$\operatorname{ri}(A) = \left\{ \sum_{i=1}^{k} \lambda_i a^i : \lambda \in (\mathbb{R}^{>0})^k, \sum_{i=1}^{k} \lambda_i = 1 \right\},$$

$$\operatorname{rb}(A) = \left\{ a \in A : \nexists \lambda \in (\mathbb{R}^{>0})^k \text{ with } \sum_{i=1}^{k} \lambda_i = 1 \text{ and } a = \sum_{i=1}^{k} \lambda_i a^i \right\}.$$

Proof Exercise 3.1 in [16]. □

For example, for a line connecting two points $a^1, a^2 \in \mathbb{R}^n$, $a^1 \neq a^2$, the relative boundary is the set containing a^1 and a^2 and the relative interior is the line without the end points, i.e., the set $\{\lambda a^1 + (1-\lambda) a^2 : \lambda \in (0,1)\}$.

References

[1] S. Scholtes, *Introduction to Piecewise Differentiable Equations*. Springer Briefs in Optimization (New York, NY : Springer New York, 2012). https://doi.org/10.1007/978-1-4614-4340-7

[2] M. Ulbrich, Nonsmooth newton-like methods for variational inequalities and constrained optimization problems in function spaces. Habilitation thesis, Fakultät für Mathematik, Technische Universität München, München, Germany (2001)

[3] F.H. Clarke, *Optimization and Nonsmooth Analysis* (Society for Industrial and Applied Mathematics, 1990). https://doi.org/10.1137/1.9781611971309

[4] A. Bagirov, N. Karmitsa, M.M. Mäkelä, *Introduction to Nonsmooth Optimization* (Springer International Publishing, 2014). https://doi.org/10.1007/978-3-319-08114-4

[5] C. Lemaréchal, in *Handbooks in Operations Research and Management Science* (Elsevier, 1989), pp. 529–572. https://doi.org/10.1016/s0927-0507(89)01008-x

[6] K. Miettinen, *Nonlinear Multiobjective Optimization* (Springer US, 1998). https://doi.org/10.1007/978-1-4615-5563-6

[7] M. Ehrgott, *Multicriteria Optimization* (Springer-Verlag, 2005). https://doi.org/10.1007/3-540-27659-9

[8] M.M. Mäkelä, V.P. Eronen, N. Karmitsa, On nonsmooth multiobjective optimality conditions with generalized convexities. Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos pp. 333–357 (2014). https://doi.org/10.1007/978-1-4939-0808-0_17

[9] M.M. Mäkelä, N. Karmitsa, O. Wilppu, Multiobjective proximal bundle method for nonsmooth optimization. TUCS technical report No 1120, Turku Centre for Computer Science, Turku (2014)

[10] B. Gebken, S. Peitz, An Efficient Descent Method for Locally Lipschitz Multiobjective Optimization Problems. Journal of Optimization Theory and Applications **80**, 3–29 (2021). https://doi.org/10.1007/s10957-020-01803-w

[11] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation **6**(2), 182–197 (2002). https://doi.org/10.1109/4235.996017

[12] C. Aliprantis, K. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, 3rd edn. (Springer-Verlag Berlin Heidelberg, 2006). https://doi.org/10.1007/3-540-29587-9

[13] R.T. Rockafellar, *Convex Analysis* (Princeton University Press, 1970). https://doi.org/10.1515/9781400873173

[14] J. Gallier, *Geometric Methods and Applications* (Springer New York, 2011). https://doi.org/10.1007/978-1-4419-9961-0

[15] D. Jungnickel, *Optimierungsmethoden* (Springer Berlin Heidelberg, 2015). https://doi.org/10.1007/978-3-642-54821-5

[16] A. Brøndsted, *An Introduction to Convex Polytopes* (Springer New York, 1983). https://doi.org/10.1007/978-1-4612-1148-8