Gebken, Bennet; Bieker, Katharina; Peitz, Sebastian

On the structure of regularization paths for piecewise differentiable regularization terms

Journal of Global Optimization

Provided in Cooperation with:
Springer Nature

Suggested Citation: Gebken, Bennet; Bieker, Katharina; Peitz, Sebastian (2022) : On the structure of regularization paths for piecewise differentiable regularization terms, Journal of Global Optimization, ISSN 1573-2916, Springer US, New York, NY, Vol. 85, Iss. 3, pp. 709-741, https://doi.org/10.1007/s10898-022-01223-2

Journal of Global Optimization (2023) 85:709–741
https://doi.org/10.1007/s10898-022-01223-2

On the structure of regularization paths for piecewise differentiable regularization terms

Bennet Gebken[1] · **Katharina Bieker**[1] · **Sebastian Peitz**[2]

Received: 11 January 2022 / Accepted: 8 August 2022 / Published online: 1 September 2022

Abstract
Regularization is used in many different areas of optimization when solutions are sought which not only minimize a given function, but also possess a certain degree of regularity. Popular applications are image denoising, sparse regression and machine learning. Since the choice of the regularization parameter is crucial but often difficult, path-following methods are used to approximate the entire regularization path, i.e., the set of all possible solutions for all regularization parameters. Due to their nature, the development of these methods requires structural results about the regularization path. The goal of this article is to derive these results for the case of a smooth objective function which is penalized by a piecewise differentiable regularization term. We do this by treating regularization as a multiobjective optimization problem. Our results suggest that even in this general case, the regularization path is piecewise smooth. Moreover, our theory allows for a classification of the nonsmooth features that occur in between smooth parts. This is demonstrated in two applications, namely support-vector machines and exact penalty methods.

Keywords Regularization · Nonsmooth analysis · Multiobjective optimization

Mathematics Subject Classification 65F22 · 62J07 · 90C29 · 49J52

1 Introduction

In optimization, *regularization* is one of the basic tools for dealing with irregular solutions. For an objective function $f : \mathbb{R}^n \to \mathbb{R}$, the idea is to add a *regularization term* $g : \mathbb{R}^n \to \mathbb{R}$ to f which enforces regularity, and to weight g with a *regularization parameter* $\lambda \geq 0$ to

✉ Bennet Gebken
bgebken@math.upb.de

Katharina Bieker
bieker@math.upb.de

Sebastian Peitz
sebastian.peitz@upb.de

[1] Department of Mathematics, Paderborn University, 33098 Paderborn, Germany

[2] Department of Computer Science, Paderborn University, 33098 Paderborn, Germany

control to which extent this regularity is enforced. So instead of optimizing f, the regularized problem

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda g(x)$$

with $\lambda \geq 0$ is solved. For $\lambda = 0$ the original problem is recovered. Increasing λ leads to successively more regular solutions, at the cost of an increased objective value of f.

Depending on the application, the term "regularity" above can have many different meanings: In sparse regression, regularity of the solution means sparsity, and a prominent example for the regularization term is the ℓ^1-norm [1, 2]. In hyperplane separation for data classification (also known as *support-vector machines*), regularity is related to robustness of the derived classifier, and a possible regularization term can be derived from the scalar product of the data points with the hyperplane (known as the *hinge loss*) [2, 3]. In image denoising, regularity means the absence of noise in the reconstructed image, which can be measured using the total variation [4]. In (exact) penalty methods for constrained optimization problems, regularity refers to feasibility, and the sum of the individual constraint violations can be used as a regularization term [5, 6]. Finally, in deep learning, regularization is used to avoid overfitting, which is usually related to the ℓ^2- or ℓ^1-norm of the weights [3, 7].

Clearly, the choice of the regularization parameter λ has a large impact on the solution of the regularized problem. If λ is chosen too small, then solutions are almost optimal for f but irregular. If it is chosen too large, then solutions are highly regular but have an unacceptably large objective value with respect to f. One way of dealing with this issue is to not only compute a regularized solution for a single λ, but to compute the entire so-called *regularization path* R, which is the set of all regularized solutions for all $\lambda \geq 0$. The properties and features of R (e.g., *knee points* [8]) can then be used to better choose a desirable solution. Obviously, simply solving the regularized problem for many $\lambda \geq 0$ to obtain a discretization of R is inefficient. Instead, so-called *path-following methods* (also known as *continuation methods*, *homotopy methods* or *predictor–corrector methods*) can be used, which iteratively compute new points on the regularization path close to already known points until the complete path is explored. By exploiting the smoothness properties of the path, the computation of each new point tends to be cheap. For the development of such methods, it is crucial to have a good understanding of the structure of the regularization path. In [9, 10], it was shown that for sparse regression, the regularization path R is piecewise linear and a path-following method was proposed for its computation. Similar results were shown in [11] for support-vector machines. In a more general setting in [12], it was shown that if f is piecewise quadratic and g is piecewise linear, then R is always piecewise linear. In case of the exact penalty method in constrained optimization, it was shown in [13] that if the constrained problem is convex (and the equality constraints are affinely linear), then R is piecewise smooth. Recently, in [14], the structure of the regularization path was analyzed for the case where f is twice continuously differentiable and g is the ℓ^1-norm, with the results suggesting that R is piecewise smooth.

The goal of this article is to analyze the structure of the regularization path in a more general setting. Note that in the applications above, we have the pattern that f is always smooth while g is always nonsmooth. Thus, in this article, we will also assume that f is smooth. For g, we will assume that it is merely *piecewise differentiable* (as defined in [15]). Compared to weaker assumptions in nonsmooth analysis like local Lipschitz continuity, this has the advantage that the Clarke subdifferential of g is easy to compute and that the set of nonsmooth points of g can essentially be described as a level set of certain smooth functions. Since all of the regularization terms in the above applications (except for the ℓ^2-norm) are in fact piecewise differentiable, our setting generalizes many of the existing approaches. We

will analyze the structure of R by approximating it with the *critical regularization path* R_c, which is based on the first-order optimality conditions of the regularized problem, and then identifying sufficient conditions for R_c to be smooth around a given point. More precisely, our main result will be that if these conditions are met, then R_c is locally the projection of a higher-dimensional smooth manifold onto $\mathbb{R}^n$ (cf. Theorem 2). In particular, all points violating these conditions are potential "kinks" (or "nonsmooth points") of R_c. Depending on which condition is violated, this allows for a classification of nonsmooth features of the regularization path. Furthermore, the nature of our sufficient conditions suggests that R_c (and R) is still piecewise smooth.

From a theoretical point of view, the core idea of this article is the application of the *level set theorem* (cf. [16], Theorem 5.12) to a smooth function h whose projected zero level set locally coincides with R_c. For h to be smooth, we have to carefully construct it by considering the so-called smooth *selection functions* that g consists of. Compared to the previous results in [9–13], this general approach has to be followed since, apart from smoothness, no other properties of the selection functions can be exploited. For the case where g is the ℓ^1-norm, this methodology reduces to the approach in [14], which is significantly easier to handle due to the simplicity of the ℓ^1-norm. In the more general case that is considered in this article, more care has to be taken when working with the selection functions.

The remainder of this article is structured as follows. In Sect. 2, we begin by briefly introducing the basic concepts that we use in our theoretical results (A more detailed introduction can be found in the electronic supplementary material). Besides piecewise differentiability, these are *multiobjective optimization* and *affine geometry*. The former can be used to obtain an (almost) equivalent formulation of the regularization problem as a multiobjective optimization problem, while the latter is required for working with the subdifferential of g. In Sect. 3, we will analyze the structure of the regularization path R. We will do this by expressing R_c as the union of the intersection of certain sets, whose structure we can analyze by applying standard results from differential geometry. We will also formulate an abstract algorithm for a path-following method based on our results. In Sect. 4, we will apply our results to two problem classes, which are support-vector machines and the exact penalty method. Finally, we draw a conclusion and discuss possible future work in Sect. 5.

2 Basic concepts

In this section, we will introduce the basic ideas of piecewise differentiable functions, multiobjective optimization and affine geometry. As these topics may not be common in the optimization community, we also compiled a more detailed introduction with the specific results that we use in this article and included it in the electronic supplementary material.

For the regularization term we will assume *piecewise differentiability* [15] in the following sense.

Definition 1 Let $U \subseteq \mathbb{R}^n$ be open. 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.

If $g : U \to \mathbb{R}$ is a PC^r-function with selection functions $\{g_1, \ldots, g_k\}$, then the Clarke subdifferential [17] of g is given by

$$\partial g(x) = \operatorname{conv}(\{\nabla g_i(x) : i \in I^e(x)\}) \quad \forall x \in U, \tag{1}$$

where conv$(\cdot)$ denotes the convex hull and $I^e(x)$ is the set of *essentially active selection functions in* x. In particular, $\partial g(x)$ is a polytope and, assuming the essentially active selection functions are known, easy to compute.

For the derivation of our theoretical results, we will interpret regularization problems as *multiobjective optimization problems* (MOPs) [18–20]. For general functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$, the MOP minimizing f and g is denoted by

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

and its solution is defined in the following.

Definition 2 A point $x \in \mathbb{R}^n$ is called *Pareto optimal* if there is no $y \in \mathbb{R}^n$ with

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

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*.

If both f and g are at least locally Lipschitz continuous and x is Pareto optimal, then

$$0 \in \text{conv}(\partial f(x) \cup \partial g(x)) \tag{2}$$

or, equivalently,

$$\exists \alpha_1, \alpha_2 \geq 0, \xi^1 \in \partial f(x), \xi^2 \in \partial g(x) : \alpha_1 \xi^1 + \alpha_2 \xi^2 = 0, \alpha_1 + \alpha_2 = 1. \tag{3}$$

Points that satisfy this optimality condition are called *Pareto critical* and the set of all such points is the *Pareto critical set* P_c. The quantities α_1 and α_2 in (3) will be referred to as *KKT multipliers of* f and g *in* x, respectively.

Finally, the structure of the condition (3) will make it possible to use *affine geometry* [21–23] to relate properties of the Pareto critical set P_c to properties of the subdifferentials $\partial f(x)$ and $\partial g(x)$.

Definition 3 (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 aff(E) is the set of all affine combinations of elements of E, called the *affine hull of* E. Formally,

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

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

Analogously to linear algebra, it is possible to define the *affine dimension* affdim(A) and *affine bases* of an affine space A. An important result about affine spaces (and convex sets) is *Carathéodory's theorem*:

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

3 The structure of the regularization path

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $g : \mathbb{R}^n \to \mathbb{R}$ be PC^1. For a *regularization parameter* $\lambda \geq 0$, consider the parameter-dependent problem

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda g(x). \tag{4}$$

The set

$$R := \left\{ \bar{x} \in \mathbb{R}^n : \exists \lambda \geq 0 \text{ with } \bar{x} \in \arg\min_{x \in \mathbb{R}^n} f(x) + \lambda g(x) \right\} \tag{5}$$

is known as the *regularization path* of (4) [11, 24, 25] and the goal of this article is to analyze its structure.

We will do this by not analyzing R directly, but by analyzing the (potentially larger) set that is defined by the first-order optimality condition of (4): If $\bar{x}$ is a solution of (4) for some $\lambda \geq 0$, then it is a *critical point* of $f + \lambda g$, i.e., $0 \in \partial(f + \lambda g)(\bar{x})$ (cf. Theorem 4.1 in [6]). This is the motivation for defining the *critical regularization path*

$$R_c := \left\{ \bar{x} \in \mathbb{R}^n : \exists \lambda \geq 0 \text{ with } 0 \in \partial(f + \lambda g)(\bar{x}) \right\}. \tag{6}$$

In general we have $R \subseteq R_c$. If $f + \lambda g$ is convex (e.g., if both f and g are convex), then criticality is sufficient for optimality (cf. Theorem 4.2 in [6]), so $R = R_c$. For example, this is the case for the Lasso problem [1] (where f contains some least squares error and g is the ℓ^1-norm) and total variation denoising [4] (where f contains some least squares error and g is the total variation). The extend to which structural result about R_c apply to R in the general nonconvex case will be discussed in Remark 5.

Our main result in this section will be that R_c has a piecewise smooth structure. More precisely, we will derive five conditions (Assumptions A1 to A5) for a point $x^0 \in R_c$ which, when combined, assure that locally around x^0, R_c is the projection of a smooth manifold from a higher-dimensional space onto $\mathbb{R}^n$. In turn, these assumptions allow for a classification of kinks of R_c by checking which assumption is violated. Throughout this article, we will use the term *kinks* to loosely refer to points in R_c around which R_c is not a smooth manifold.

In order to analyze the structure of R_c, we first show that R_c is related to the Pareto critical set P_c of the MOP

$$\min_{x \in \mathbb{R}^n} \begin{pmatrix} f(x) \\ g(x) \end{pmatrix}. \tag{7}$$

More precisely, we have the following lemma.

Lemma 1 *It holds:*

(a) $R_c = \{\bar{x} \in \mathbb{R}^n : \exists \xi \in \partial g(\bar{x}), \alpha_1 > 0, \alpha_2 \geq 0 \text{ with } \alpha_1 \nabla f(\bar{x}) + \alpha_2 \xi = 0 \text{ and } \alpha_1 + \alpha_2 = 1\} \subseteq P_c$.
(b) $R_c \cup \{x \in \mathbb{R}^n : 0 \in \partial g(x)\} = P_c$.

Proof (a) Since f is continuously differentiable we have $\partial f(x) = \{\nabla f(x)\}$ for all $x \in \mathbb{R}^n$. Furthermore, from basic calculus for subdifferentials (cf. Corollary 1 in [17], Section

2.3) it follows that $\bar{x} \in R_c$ is equivalent to

$$\begin{aligned} &\exists \lambda \geq 0 : 0 \in \partial(f + \lambda g)(\bar{x}) = \partial f(\bar{x}) + \lambda \partial g(\bar{x}) = \nabla f(\bar{x}) + \lambda \partial g(\bar{x}) \\ &\Leftrightarrow \ \exists \lambda \geq 0 : 0 \in \frac{1}{1+\lambda} \nabla f(\bar{x}) + \frac{\lambda}{1+\lambda} \partial g(\bar{x}) \\ &\Leftrightarrow \ \exists \xi \in \partial g(\bar{x}), \lambda \geq 0 : \frac{1}{1+\lambda} \nabla f(\bar{x}) + \frac{\lambda}{1+\lambda} \xi = 0 \\ &\Leftrightarrow \ \exists \xi \in \partial g(\bar{x}), \alpha_1 > 0, \alpha_2 \geq 0 : \alpha_1 \nabla f(\bar{x}) + \alpha_2 \xi = 0 \text{ and } \alpha_1 + \alpha_2 = 1. \end{aligned} \tag{8}$$

By (3) this implies $\bar{x} \in P_c$.

(b) Due to (a) we only have to show the implication "$\supseteq$", so let $\bar{x} \in P_c$. By (3) there are $\xi \in \partial g(\bar{x})$ and $\alpha_1, \alpha_2 \geq 0$ with $\alpha_1 + \alpha_2 = 1$ and $\alpha_1 \nabla f(\bar{x}) + \alpha_2 \xi = 0$. If $\alpha_1 = 0$ then $\alpha_2 = 1$, so $0 = \xi \in \partial g(\bar{x})$. Otherwise, $\alpha_1 > 0$ and from (8) it follows that $\bar{x} \in R_c$ (with $\lambda = \frac{\alpha_2}{\alpha_1}$). □

By the previous lemma, R_c and P_c coincide up to critical points of g in which all KKT multipliers corresponding to f are zero. Roughly speaking, these points correspond to "$\lambda = \infty$" in (4).

Remark 1 It is important to note that Lemma 1 does not imply that critical points of g are not contained in R_c, i.e., that $R_c \cap \{x \in \mathbb{R}^n : 0 \in \partial g(x)\} = \emptyset$. For example, if $0 \in \text{int}(\partial g(x))$, then it is possible to show that there is some $\bar{\lambda}$ with $0 \in \partial(f + \lambda g)(x)$ for all $\lambda \geq \bar{\lambda}$.

By Lemma 1, structural results about Pareto critical sets can be used to analyze the structure of the critical regularization path R_c. For example, under some mild regularity assumptions on f and g, Theorem 5.1 in [26] shows that in areas where g is (twice continuously) differentiable, the set of Pareto critical points with non-vanishing KKT multipliers is the projection of a 1-dimensional manifold from $\mathbb{R}^{n+2}$ onto $\mathbb{R}^n$. To derive our main result, we will extend the ideas in [26] to the whole Pareto critical set up to certain kinks.

We begin by taking a closer look at the Pareto critical set P_c of (7). By definition, P_c is characterized by the optimality condition (2). Since f is continuously differentiable and g is PC^1, the subdifferential of f is simply its gradient, and the subdifferential of g is the convex hull of all essentially active selection functions (cf. (1)). Thus, for a fixed $x \in \mathbb{R}^n$, (2) is equivalent to the existence of a vanishing convex combination of a finite number of elements. This is the same type of condition as in the smooth case, except that there is now no continuous dependency of these elements on x. Furthermore, the number of elements is not constant. Nonetheless, by iterating over all possible essentially active sets, P_c can at least be written as the union of sets that behave similarly to Pareto critical sets in the smooth case. Let $\{g_1, \ldots, g_k\}$ be a set of selection functions of g. Then formally, these considerations lead to the following decomposition of P_c:

$$\begin{aligned} P_c &= \{x \in \mathbb{R}^n : 0 \in \text{conv}(\{\nabla f(x)\} \cup \partial g(x))\} \\ &= \{x \in \mathbb{R}^n : 0 \in \text{conv}(\{\nabla f(x)\} \cup \{\nabla g_i(x) : i \in I^e(x)\})\} \\ &= \bigcup_{I \subseteq \{1,\ldots,k\}} P_c^I \cap \Omega^I, \end{aligned} \tag{9}$$

where

$$\begin{aligned} P_c^I &:= \{x \in \mathbb{R}^n : 0 \in \text{conv}(\{\nabla f(x)\} \cup \{\nabla g_i(x) : i \in I\})\}, \\ \Omega^I &:= \{x \in \mathbb{R}^n : I^e(x) = I\}. \end{aligned} \tag{10}$$

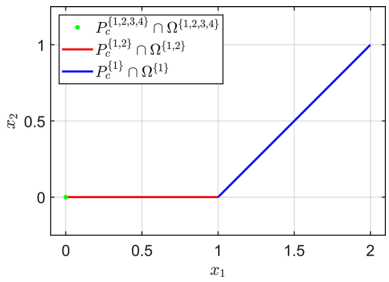

Fig. 1 Decomposition of P_c into the sets $P_c^I \cap \Omega^I$ as in (9)

In words, P_c^I is the Pareto critical set of the (smooth) MOP with objective vector $(f, g_{i_1}, \dots g_{i_{|I|}})^\top$ (for $I = \{i_1, \dots, i_{|I|}\}$) and Ω^I is the set of points in $\mathbb{R}^n$ in which precisely the selection functions with an index in I are essentially active. Thus, (9) expresses P_c as the union of Pareto critical sets of smooth MOPs that are intersected with the sets of points with constant essentially active sets. A visualization of this decomposition is shown in the following example.

Example 1 Consider problem (7) for $f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto (x_1 - 2)^2 + (x_2 - 1)^2$, and

$$\begin{aligned}
g_1 &: \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto x_1 + x_2,\\
g_2 &: \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto x_1 - x_2,\\
g_3 &: \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto -x_1 + x_2,\\
g_4 &: \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto -x_1 - x_2,\\
g &: \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto \max(\{g_1(x), g_2(x), g_3(x), g_4(x)\}) = \|x\|_1.
\end{aligned}$$

It is possible to show that the Pareto critical (and in this case Pareto optimal) set is given by

$$\begin{aligned}
P_c &= \{(0,0)^\top\} \cup ((0,1] \times \{0\}) \cup \{x \in \mathbb{R}^2 : x_1 \in (1,2], x_2 = x_1 - 1\}\\
&= (P_c^{\{1,2,3,4\}} \cap \Omega^{\{1,2,3,4\}}) \cup (P_c^{\{1,2\}} \cap \Omega^{\{1,2\}}) \cup (P_c^{\{1\}} \cap \Omega^{\{1\}}).
\end{aligned}$$

Figure 1 shows the decomposition of P_c into the sets $P_c^I \cap \Omega^I$ as in (9).

We will analyze the piecewise smooth structure of P_c via (9) by first analyzing Ω^I, then the intersection $P_c^I \cap \Omega^I$ and finally the union over all $P_c^I \cap \Omega^I$. Furthermore, as we expect P_c to possess kinks, we will only consider its local structure around a given point. In other words, for $x^0 \in P_c$, we will only consider the structure of $P_c \cap U$ for open neighborhoods $U \subseteq \mathbb{R}^n$ of x^0.

The strategy for our analysis in this section is to derive assumptions for x^0 which are sufficient for P_c to have a smooth structure locally around x^0. These assumptions represent different sources and types of nonsmoothness of P_c and will allow for a classification of nonsmooth points.

3.1 The structure of Ω^I

By definition, the set Ω^I only depends on g. For $I = \{i\} \subseteq \{1, \dots, k\}$, $\Omega^{\{i\}}$ is the set of points where only the selection function g_i is essentially active. From Lemma SM1 it follows that $\Omega^{\{i\}}$ is an open subset of $\mathbb{R}^n$ in this case. For $I \subseteq \{1, \dots, k\}$ with $|I| > 1$, Ω^I is the set of points where precisely the selection functions corresponding to the elements of I are

essentially active. Typically (but not necessarily), these are points where g is nonsmooth, which by Rademacher's Theorem ([27], Theorem 3.2) form a null set. In the following, we will analyze its structure.

Since we are only interested in the structure of Ω^I in a local sense, we also only have to consider restrictions $g|_U$ of g to open neighborhoods of a point $x^0 \in \mathbb{R}^n$. In terms of the open neighborhood U of x^0 and the set of selection functions of $g|_U$, we introduce the following assumption:

Assumption A1 For $x^0 \in \mathbb{R}^n$ there is an open neighborhood $U \subseteq \mathbb{R}^n$ of x^0 and a set of selection functions $\{g_1, \ldots, g_k\}$ of $g|_U$ such that

(i) $I(x^0) = \{1, \ldots, k\}$,
(ii) $I^e(x) = I(x) \quad \forall x \in U$,
(iii) $\operatorname{affdim}(\operatorname{aff}(\{\nabla g_i(x) : i \in \{1, \ldots, k\}\})) = \operatorname{affdim}(\operatorname{aff}(\{\nabla g_i(x^0) : i \in \{1, \ldots, k\}\})) \; \forall x \in U$.

Assumption A1 can be interpreted as follows: A1(i) ensures that all selection functions we consider are actually relevant for the representation of g in U. The condition A1(ii) ensures that it does not matter if we consider the active or the essentially active set in U, which allows for an easier representation of Ω^I. Finally, A1(iii) makes sure that the representation of $\partial g(x^0)$ via the gradients of our selection functions is "stable" on U with respect to its affine dimension.

In the following, we will discuss the restrictiveness of Assumption A1. By (SM1), A1(i) can always be satisfied by choosing U sufficiently small. For A1(ii) and (iii), we consider the following example.

Example 2 (a) Let

$$\begin{aligned} g_1 &: \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto x_2^2 - x_1, \\ g_2 &: \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto \begin{cases} x_1^2 - x_1, & x_1 \leq 0, \\ -x_1, & x_1 > 0, \end{cases} \\ g &: \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto \max(\{g_1(x), g_2(x)\}). \end{aligned}$$

Then g is PC^1 with selection functions g_1 and g_2. The graph and the level sets of g are shown in Fig. 2. For the activity of g_2 we have

$$2 \in I(x) \;\Leftrightarrow\; g(x) = g_2(x) \;\Leftrightarrow\; \begin{cases} x_2 \in [x_1, -x_1], & x_1 \leq 0, \\ x_2 = 0, & x_1 > 0, \end{cases}$$

and

$$\begin{aligned} 2 \in I^e(x) &\;\Leftrightarrow\; x \in \operatorname{cl}(\operatorname{int}(\{y \in \mathbb{R}^2 : g(y) = g_2(y)\})) \\ &\;\Leftrightarrow\; x_1 \leq 0, \; x_2 \in [x_1, -x_1]. \end{aligned}$$

Thus, for any open neighborhood $U \subseteq \mathbb{R}^2$ of $x^0 = (0, 0)^\top$, there is some $x \in U$ with $I^e(x) \neq I(x)$. In other words, A1(ii) does not hold in x^0 for this set of selection functions. But note that in this case, this can easily be fixed by modifying the behavior of g_2 for $x_1 > 0$. For example, replacing g_2 by

$$\tilde{g}_2 : \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto \begin{cases} x_1^2 - x_1, & x_1 \leq 0, \\ -x_1^2 - x_1, & x_1 > 0. \end{cases}$$

solves the issue.

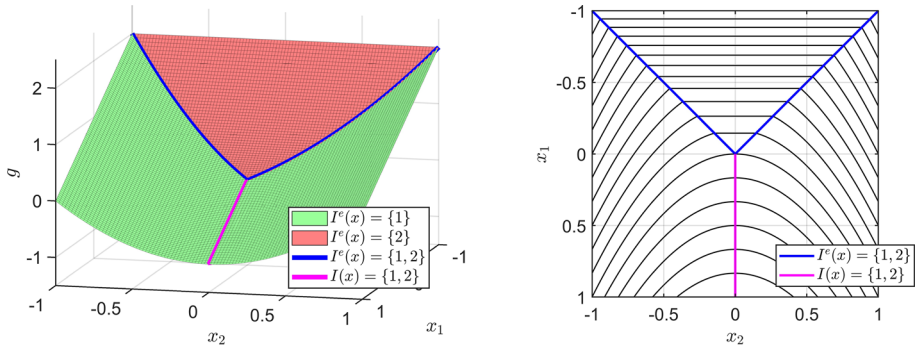

Fig. 2 **a** The graph of the PC^1-function g in Example 2a. **b** The level sets of g

(b) For the selection functions g_1 and $\tilde{g}_2$ of g as in a), we have

$$\nabla g_1(x) = \begin{pmatrix} -1 \\ 2x_2 \end{pmatrix} \text{ and } \nabla \tilde{g}_2(x) = \begin{cases} (2x_1 - 1, 0)^\top, & x_1 \leq 0, \\ (-2x_1 - 1, 0)^\top, & x_1 > 0. \end{cases}$$

In particular, in $x^0 = (0, 0)^\top$ we have $\nabla g_1(x^0) = \nabla \tilde{g}_2(x^0) = (-1, 0)^\top$, so

$$\operatorname{affdim}(\operatorname{aff}(\{\nabla g_1(x^0), \nabla \tilde{g}_2(x^0)\})) = 0.$$

But it is easy to see that

$$\operatorname{affdim}(\operatorname{aff}(\{\nabla g_1(x), \nabla \tilde{g}_2(x)\})) = 1 \quad \forall x \in \mathbb{R}^2 \setminus \{0\}.$$

In particular, A1(iii) does not hold in x^0 (for this set of selection functions).

By Lemma SM1, for a given $x^0 \in \mathbb{R}^n$, we can always choose the open neighborhood U of x^0 such that all selection functions of the local restriction $g|_U$ of g are essentially active in x^0. In particular, we can assume that $I^e(x^0) = I(x^0)$. While this does not imply that (ii) holds in Assumption A1, the previous example shows how A1(ii) may be satisfied through modifications of the selection functions in areas where they are active, but not essentially active. Although we will not prove that this is always possible, it motivates us to believe that A1(ii) is not a strong assumption in practice.

In contrast to A1(ii), modifying the selection functions will have less impact on A1(iii). The reason for this is the fact that if A1(i) and A1(ii) hold, then the right-hand side of A1(iii) is the dimension of the affine hull of the subdifferential of g in x^0 (cf. (1)). In particular, the right-hand side does not depend on the choice of selection functions. In light of this, A1(iii) implies that the dimension of the affine hull of the subdifferential of g is constant in all $x \in U$ with $I^e(x) = I^e(x^0)$, i.e., in all $x \in \Omega^{I^e(x^0)}$ (cf. (10)). Thus, A1(iii) is more related to the function g and less related to the choice of selection functions. In Example 2 a), we see that the set $\Omega^{\{1,2\}}$ (in blue) has a kink in $x^0 = (0, 0)^\top$. The following lemma suggests that this is caused by A1(iii) being violated. Thus, by assuming A1(iii), we limit ourselves to local restrictions $g|_U$ for which $\Omega^{I^e(x^0)}$ has a smooth structure.

Lemma 2 *Let $x^0 \in \mathbb{R}^n$. Let $U \subseteq \mathbb{R}^n$ be an open neighborhood of x^0 and let $\{g_1, \ldots, g_k\}$ be a set of selection functions of $g|_U$ as in Assumption A1. Let $d = \operatorname{affdim}(\operatorname{aff}(\partial g(x^0)))$ and*

let $\{i_1, \ldots, i_{d+1}\} \subseteq \{1, \ldots, k\}$ *such that* $\{\nabla g_i(x^0) : i \in \{i_1, \ldots, i_{d+1}\}\}$ *is an affine basis of* $\operatorname{aff}(\{\nabla g_i(x^0) : i \in \{1, \ldots, k\}\})$. *Then there is an open neighborhood* $U' \subseteq U$ *of* x^0 *such that*

$$g_i(x) - g_1(x) = 0 \; \forall i \in \{2, \ldots, k\} \quad \Leftrightarrow \quad g_i(x) - g_{i_1}(x) = 0 \; \forall i \in \{i_2, \ldots, i_{d+1}\}$$

for all $x \in U'$ *and* $\Omega^{\{1,\ldots,k\}} \cap U'$ *is an embedded* $(n-d)$*-dimensional submanifold of* U'. *In particular,*

$$\Omega^{\{1,\ldots,k\}} \cap U' = \{x \in U' : g_i(x) - g_{i_1}(x) = 0 \; \forall i \in \{i_2, \ldots, i_{d+1}\}\}.$$

Proof The direction "⇒" is obvious, so consider the converse. By A1(iii) and since the gradients $\nabla g_i, i \in \{i_1, \ldots, i_{d+1}\}$, are continuous, there is an open neighborhood $U' \subseteq U$ of x^0 such that $\{\nabla g_i(x) : i \in \{i_1, \ldots, i_{d+1}\}\}$ is an affine basis of $\{\nabla g_i(x) : i \in \{1, \ldots, k\}\}$ for all $x \in U'$. Let

$$\varphi : U' \to \mathbb{R}^{k-1}, \quad x \mapsto \begin{pmatrix} g_2(x) - g_1(x) \\ \vdots \\ g_k(x) - g_1(x) \end{pmatrix}.$$

By A1(iii) the Jacobian $D\varphi(x)$ has constant rank d for all $x \in U'$. By A1(i) we have $\varphi(x^0) = 0$, so the level set $L := \varphi^{-1}(0) = \Omega^{\{1,\ldots,k\}} \cap U'$ is nonempty. Thus, by Theorem 5.12 in [16], L is an embedded $(n-d)$-dimensional submanifold of U'. Additionally, let

$$\varphi' : U' \to \mathbb{R}^{d}, \quad x \mapsto \begin{pmatrix} g_{i_2}(x) - g_{i_1}(x) \\ \vdots \\ g_{i_{d+1}}(x) - g_{i_1}(x) \end{pmatrix}.$$

By construction, $D\varphi'(x)$ has constant rank d for all $x \in U'$. With the same argument as above, it follows that $L' := \varphi'^{-1}(0)$ is an embedded $(n-d)$-dimensional submanifold of U' as well. Since $L \subseteq L'$, L is also an embedded $(n-d)$-dimensional submanifold of L' (cf. [16], Proposition 4.22). By Proposition 5.1 in [16], this implies that L is an open subset of L'. As L' is endowed with the subspace topology of $U' \subseteq \mathbb{R}^n$, this means that we can assume w.l.o.g. that U' is an open neighborhood of x^0 with $U' \cap L' = L$, completing the proof. □

By the previous lemma, Assumption A1 allows us to assume w.l.o.g. that for the restriction $g|_U$, the set of points with a constant active set $\Omega^{I^e(x^0)}$ is a smooth manifold around $x^0 \in U$ of dimension $n - \operatorname{affdim}(\operatorname{aff}(\partial g(x^0)))$. Furthermore, it shows that for the representation of $\Omega^{I^e(x^0)}$ as a level set, it is sufficient to only consider a subset of the set of selection functions whose gradients form an affine basis of $\partial g(x^0)$.

3.2 The structure of $P_c^I \cap \Omega^I$

After analyzing the structure of Ω^I, we will now turn towards the structure of the intersection $P_c^I \cap \Omega^I$ in (9). First of all, as for Ω^I, we will show that not all selection functions of g are required for the representation of $P_c^I \cap \Omega^I$. More precisely, a simple application of Carathéodory's theorem (Theorem 1) to the definition of P_c^I yields the following result.

Lemma 3 *Let* $x^0 \in P_c$ *and let* $\{g_1, \ldots, g_k\}$ *be a set of selection functions of* g. *If* x^0 *is not a critical point of* g, *then there is an index set* $\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, k\}$ *with* $r = \operatorname{affdim}(\operatorname{aff}(\{\nabla f(x^0)\} \cup \partial g(x^0)))$ *such that*

(a) $0 \in \operatorname{conv}(\{\nabla f(x^0)\} \cup \{\nabla g_i(x^0) : i \in \{i_1, \ldots, i_r\}\})$,
(b) $\{\nabla f(x^0)\} \cup \{\nabla g_i(x^0) : i \in \{i_1, \ldots, i_r\}\}$ *is affinely independent.*

Proof By Theorem 1, there is an affinely independent subset of

$$\{\nabla f(x^0)\} \cup \{\nabla g_i(x^0) : i \in \{1, \ldots, k\}\}$$

of size $r + 1$ with zero in its convex hull. Since x^0 is not a critical point of g, $\nabla f(x^0)$ must be contained in that subset. □

With Lemma 2 and Lemma 3, we have ways to simplify Ω^I and P_c^I, respectively, by only considering certain selection functions of g. But note that we can not necessarily choose the same selection functions for both results: Although the set $\{\nabla g_i(x^0) : i \in \{i_1, \ldots, i_r\}\}$ in Lemma 3 is affinely independent, the index set $\{i_1, \ldots, i_r\}$ can not necessarily be used in Lemma 2 since we might have $r < d + 1$, i.e.,

$$\begin{aligned} &\operatorname{affdim}(\operatorname{aff}(\{\nabla f(x^0)\} \cup \partial g(x^0))) < \operatorname{affdim}(\operatorname{aff}(\partial g(x^0))) + 1 \\ \Leftrightarrow\ &\operatorname{aff}(\{\nabla f(x^0)\} \cup \partial g(x^0)) = \operatorname{aff}(\partial g(x^0)) \\ \Leftrightarrow\ &\nabla f(x^0) \in \operatorname{aff}(\partial g(x^0)). \end{aligned} \tag{11}$$

In particular, since x^0 is Pareto critical, this would imply that $0 \in \operatorname{aff}(\partial g(x^0))$ (even though x^0 is not critical for g, i.e., $0 \notin \operatorname{conv}(\partial g(x^0))$). The following lemma shows that this scenario is related to the uniqueness of the KKT multiplier corresponding to f in x^0.

Lemma 4 *Let $x^0 \in P_c$ such that x^0 is not a critical point of g.*

(a) *If the KKT multiplier α_1 of f in x^0 (cf. (3)) is not unique, then $\nabla f(x^0) \in \operatorname{aff}(\partial g(x^0))$.*
(b) *If $\nabla f(x^0) \in \operatorname{aff}(\partial g(x^0))$ and 0 is contained in the relative interior (cf. Definition SM9) of $\operatorname{conv}(\{\nabla f(x^0)\} \cup \partial g(x^0))$, then the KKT multiplier α_1 of f in x^0 is not unique.*

Proof See "Appendix A.1". □

Remark 2 In [26], Section 4.3, it was shown that in the smooth case and under certain regularity assumptions on f and g, the coefficient vector of the vanishing convex combination in the KKT condition in a point $x \in P_c$, i.e., the vector $(\alpha_1, \alpha_2)^\top$ in (3), is orthogonal to the tangent space of the image of the Pareto critical set at $(f(x), g(x))^\top$. Thus, roughly speaking, non-uniqueness of $(\alpha_1, \alpha_2)^\top$ suggests that this tangent space is "degenarate", i.e., that the Pareto front possesses a kink at $(f(x), g(x))^\top$.

The following example shows what behavior may occur if the KKT multiplier of f is not unique.

Example 3 Consider problem (7) for $f : \mathbb{R}^2 \to \mathbb{R}$, $x \mapsto x_1^2 + x_2^2$, and

$$\begin{aligned} g_1 &: \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto x_1^2 + (x_2 - 1)^2, \\ g_2 &: \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto x_1^2 + (x_2 - 1)^2 - \left(x_2 - \frac{1}{2}\right), \\ g &: \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto \max(\{g_1(x), g_2(x)\}). \end{aligned}$$

Then g is PC^1 with selection functions g_1 and g_2. It is easy to see that

$$\Omega^{\{1,2\}} = \{x \in \mathbb{R}^n : I^e(x) = \{1, 2\}\} = \mathbb{R} \times \left\{\frac{1}{2}\right\},$$

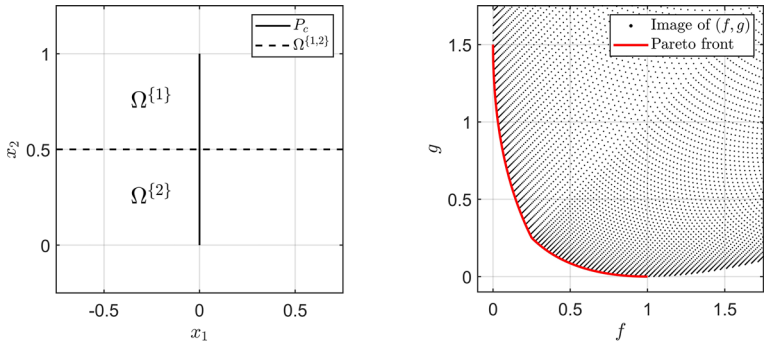

Fig. 3 **a** Pareto critical set P_c and Ω^I, $I \subseteq \{1, 2\}$, in Example 3. **b** Pointwise discretization of the image $\{(f(x), g(x))^\top : x \in \mathbb{R}^2\}$ of the objective vector (f, g) and the image of the Pareto critical set under (f, g)

as depicted in Fig. 3a.

The Pareto critical (and in this case Pareto optimal) set is given by $P_c = \{0\} \times [0, 1]$. In particular, $x^0 = (0, \frac{1}{2})^\top$ is the only Pareto critical point where more than one selection function is active, i.e., $P_c^{\{1,2\}} \cap \Omega^{\{1,2\}} = \{x^0\}$. By computing the gradients in x^0, we obtain

$$\nabla f(x^0) = (0, 1)^\top, \ \nabla g_1(x^0) = (0, -1)^\top, \ \nabla g_2(x^0) = (0, -2)^\top.$$

We see that

$$\frac{1}{2}\nabla f(x^0) + \frac{1}{2}\nabla g_1(x^0) = 0 \quad \text{and} \quad \frac{2}{3}\nabla f(x^0) + \frac{1}{3}\nabla g_2(x^0) = 0,$$

so the KKT multiplier of f is not unique. By Lemma 4 this implies $\nabla f(x^0) \in \operatorname{aff}(\{\partial g(x^0)\})$. More explicitly, for this example, it is easy to check that

$$\nabla f(x^0) = 3\nabla g_1(x^0) - 2\nabla g_2(x^0).$$

Figure 3b shows an approximation of the image of (f, g) and the image of the Pareto critical set. As discussed in Remark 2, we see that the image of P_c has a kink at $(f(x^0), g(x^0))^\top = (\frac{1}{4}, \frac{1}{4})^\top$.

As the previous example suggests, a scenario where the KKT multiplier of f is not unique may occur if the Pareto critical set goes transversally through the set of nonsmooth points instead of moving tangentially along it. In other words, it may occur if arbitrarily close to $x^0 \in P_c$, there are Pareto critical points with essentially active sets I_1 and I_2 such that $I_1 \neq I_2$ and $I_1 \neq I^e(x^0) \neq I_2$. Due to continuity of the gradients, the KKT multipliers for both sets I_1 and I_2 have accumulation points that are KKT multipliers of x^0. Since $I_1 \neq I_2$, these accumulation points may not coincide, such that the KKT multipliers in x^0 are not unique. In terms of the structure of $P_c^I \cap \Omega^I$, we see that it is a 0-dimensional set in Example 3 (for $I = \{1, 2\}$) as it is just a single point.

Although Pareto critical points x^0 with $\nabla f(x^0) \in \operatorname{aff}(\partial g(x^0))$ may not necessarily cause nonsmoothness of P_c, we will still exclude them from our consideration of the local structure of P_c around x^0 to avoid the irregularities discussed above. So formally, we introduce the following assumption:

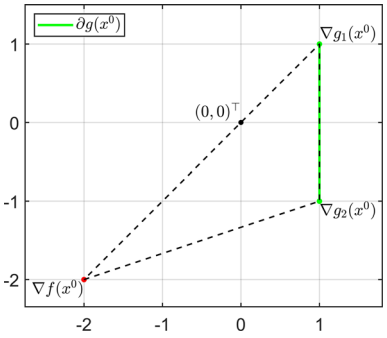

Fig. 4 The gradients of f, g_1 and g_2 in $x^0 = (1, 0)^\top$ in Example 4. The dashed line shows the (relative) boundary of the convex hull $\text{conv}(\{\nabla f(x^0)\} \cup \partial g(x^0))$

Assumption A2 For $x^0 \in P_c$ we have

$$\nabla f(x^0) \notin \text{aff}(\partial g(x^0)).$$

Roughly speaking, since $\text{affdim}(\text{aff}(\partial g(x^0))) < n$ in most cases, we expect that the set of points that violate Assumption A2 is small compared to P_c (or even empty). By (11), Assumption A2 implies that there is an index set as in Lemma 3 that satisfies the requirements of Lemma 2. In particular, $P_c^I \cap \Omega^I$ can then be expressed using only a subset of the selection functions of g.

The discussion of $P_c^I \cap \Omega^I$ so far was mainly focused on the removal of redundant information in the subdifferential of g to simplify our analysis. We will now turn towards its actual geometrical structure. To this end, we again consider Example 1.

Example 4 Let f and g be as in Example 1 (The corresponding Pareto critical set is shown in Fig. 1). Let $x^0 = (1, 0)^\top$ and $U \subseteq \mathbb{R}^2$ be the open ball with radius one around x^0. Then a set of selection functions of $g|_U$ is given by $\{g_1, g_2\}$ and we have $P_c^{\{1,2\}} \cap \Omega^{\{1,2\}} = (0, 1] \times \{0\}$. In particular, x^0 is a boundary point of $P_c^{\{1,2\}} \cap \Omega^{\{1,2\}}$, such that $P_c^{\{1,2\}} \cap \Omega^{\{1,2\}}$ is not smooth around x^0 (in the sense of smooth manifolds). The gradients of f, g_1 and g_2 are shown in Fig. 4.

We see that there is a unique convex combination

$$\frac{1}{3}\nabla f(x^0) + \frac{2}{3}\nabla g_1(x^0) + 0\nabla g_2(x^0) = 0 \tag{12}$$

where the coefficient of $\nabla g_2(x^0)$ is zero.

Note that in the previous example, there is still a vanishing affine combination of the gradients of f, g_1 and g_2 for $x = (x_1, 0)^\top$, $x_1 > 1$. But it is not a convex combination, as the coefficient corresponding to $\nabla g_2(x)$ is negative. Due to the continuity of the gradients, this can only happen if one of the coefficients in x^0 is already zero (as in (12)). To exclude the type of nonsmoothness caused by this, we introduce the following assumption.

Assumption A3 For $x^0 \in P_c$ and a set of selection functions $\{g_1, \ldots, g_k\}$ of g, there is an index set $\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, k\}$ as in Lemma 3 and positive coefficients $\alpha^0 > 0$, $\beta^0 \in (\mathbb{R}^{>0})^r$ with $\alpha^0 + \sum_{j=1}^r \beta_j^0 = 1$ and $\alpha^0 \nabla f(x^0) + \sum_{j=1}^r \beta_j^0 \nabla g_{i_j}(x^0) = 0$.

The following lemma yields a necessary condition for Assumption A3 to hold, which is related to the relative interior (cf. Definition SM9) of $\text{conv}(\{\nabla f(x^0)\} \cup \partial g(x^0))$. In particular, it is independent of the choice of selection functions.

Lemma 5 *Let $x^0 \in P_c$. If there is a set of selection functions such that Assumption A3 holds, then*

$$0 \in \mathrm{ri}(\mathrm{conv}(\{\nabla f(x^0)\} \cup \partial g(x^0))).$$

Proof See "Appendix A.2". □

After introducing the Assumptions A1, A2 and A3, we are now able to show the first structural result about $P_c^I \cap \Omega^I$. The following lemma shows that $P_c^I \cap \Omega^I$ is the projection of a level set from a higher-dimensional space onto the variable space $\mathbb{R}^n$.

Lemma 6 *Let $x^0 \in P_c$. Let $U \subseteq \mathbb{R}^n$ be an open neighborhood of x^0 and let $\{g_1, \ldots, g_k\}$ be a set of selection functions of $g|_U$ satisfying Assumptions A1 and A3. Assume that Assumption A2 holds. Then there is an index set $\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, k\}$ and an open neighborhood $U' \subseteq U$ of x^0 such that*

$$P_c^{\{1,\ldots,k\}} \cap \Omega^{\{1,\ldots,k\}} \cap U' = \mathrm{pr}_x(h^{-1}(0)) \cap U', \tag{13}$$

where $\mathrm{pr}_x : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^r \to \mathbb{R}^n$ is the projection onto the first n components and

$$h : \mathbb{R}^n \times \mathbb{R}^{>0} \times (\mathbb{R}^{>0})^r \to \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^{r-1}, (x, \alpha, \beta) \mapsto \begin{pmatrix} \alpha \nabla f(x) + \sum_{j=1}^r \beta_j \nabla g_{i_j}(x) \\ \alpha + \sum_{j=1}^r \beta_j - 1 \\ (g_{i_j}(x) - g_{i_1}(x))_{j \in \{2,\ldots,r\}} \end{pmatrix}.$$

Proof Let $\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, k\}$ be an index set as in A3. Since the gradients ∇f and ∇g_{i_j}, $j \in \{1, \ldots, r\}$, are continuous and $\{\nabla f(x^0)\} \cup \{\nabla g_{i_j}(x^0) : j \in \{1, \ldots, r\}\}$ is affinely independent, there is an open neighborhood $U' \subseteq U$ of x^0 such that $\{\nabla f(x)\} \cup \{\nabla g_{i_j}(x) : j \in \{1, \ldots, r\}\}$ is affinely independent for all $x \in U'$. In particular,

$$\begin{aligned} r &\leq \mathrm{affdim}(\mathrm{aff}(\{\nabla f(x)\} \cup \{\nabla g_i(x) : i \in \{1, \ldots, k\})) \\ &\leq \mathrm{affdim}(\mathrm{aff}(\{\nabla g_i(x) : i \in \{1, \ldots, k\})) + 1 \quad \forall x \in U'. \end{aligned} \tag{14}$$

By A1, A2 and A3, we have

$$\begin{aligned} r &\overset{A3}{=} \mathrm{affdim}(\mathrm{aff}(\{\nabla f(x^0)\} \cup \partial g(x^0))) \overset{A2}{=} \mathrm{affdim}(\mathrm{aff}(\partial g(x^0))) + 1 \\ &\overset{A1(i),(ii)}{=} \mathrm{affdim}(\mathrm{aff}(\{\nabla g_i(x^0) : i \in \{1, \ldots, k\})) + 1 \\ &\overset{A1(iii)}{=} \mathrm{affdim}(\mathrm{aff}(\{\nabla g_i(x) : i \in \{1, \ldots, k\})) + 1 \quad \forall x \in U'. \end{aligned} \tag{15}$$

Combining (14) and (15), we obtain

$$\mathrm{affdim}(\mathrm{aff}(\{\nabla f(x)\} \cup \{\nabla g_i(x) : i \in \{1, \ldots, k\})) = r \quad \forall x \in U',$$

so $\{\nabla f(x)\} \cup \{\nabla g_{i_j}(x) : j \in \{1, \ldots, r\}\}$ is an affine basis of $\{\nabla f(x)\} \cup \{\nabla g_i(x) : i \in \{1, \ldots, k\}\}$ for all $x \in U'$.

Let $x \in P_c^{\{1,\ldots,k\}} \cap \Omega^{\{1,\ldots,k\}} \cap U'$. By Lemma SM4, every element of $\mathrm{aff}(\{\nabla f(x)\} \cup \{\nabla g_i(x) : i \in \{1, \ldots, k\}\})$ can be uniquely written as an affine combination of elements of $\{\nabla f(x)\} \cup \{\nabla g_{i_j}(x) : j \in \{1, \ldots, r\}\}$. Let α^0 and β^0 as in A3. Since $\alpha^0 > 0$, $\beta^0 \in (\mathbb{R}^{>0})^r$ and the gradients ∇f, ∇g_{i_j}, $j \in \{1, \ldots, r\}$, are continuous, we can assume w.l.o.g. that U' is small enough such that there are $\alpha > 0$, $\beta \in (\mathbb{R}^{>0})^r$ with $\alpha + \sum_{j=1}^r \beta_j = 1$ and

$$\alpha \nabla f(x) + \sum_{j=1}^r \beta_j \nabla g_{i_j}(x) = 0.$$

Furthermore, $g_{i_j}(x) - g_{i_1}(x) = 0$ holds for all $j \in \{2, \dots, r\}$ since $x \in \Omega^{\{1,\dots,k\}}$. Thus, $h(x, \alpha, \beta) = 0$, i.e., $x \in \mathrm{pr}_x(h^{-1}(0)) \cap U'$.
Now let $x \in \mathrm{pr}_x(h^{-1}(0)) \cap U'$. Then $x \in P_c^{\{1,\dots,k\}}$ trivially holds since $\{i_1, \dots, i_r\} \subseteq \{1, \dots, k\}$. By A1 and Lemma 2, we can assume w.l.o.g. that U' is small enough such that $g_{i_j}(x) - g_{i_1}(x) = 0$ for all $j \in \{2, \dots, r\}$ implies $x \in \Omega^{\{1,\dots,k\}}$, completing the proof. □

Up to this point, we assumed f to be continuously differentiable and g to be PC^1. This means that the map h in the previous lemma is at least continuous. If h is actually continuously differentiable, then standard results from differential geometry can be used to analyze the structure of its level sets on the right-hand side of (13). To this end, we will assume for the remainder of this section that f is twice continuously differentiable and g is PC^2.

Theorem 2 *In the setting of Lemma 6 it holds:*

(a) *If $Dh(x, \alpha, \beta)$ has full rank for all $(x, \alpha, \beta) \in h^{-1}(0)$, then $h^{-1}(0)$ is a 1-dimensional submanifold of $\mathbb{R}^n \times \mathbb{R}^{>0} \times (\mathbb{R}^{>0})^r$.*
(b) *If $Dh(x, \alpha, \beta)$ has constant rank $m \in \mathbb{N}$ for all $(x, \alpha, \beta) \in \mathbb{R}^n \times \mathbb{R}^{>0} \times (\mathbb{R}^{>0})^r$, then $h^{-1}(0)$ is an $(n + r + 1 - m)$-dimensional submanifold of $\mathbb{R}^n \times \mathbb{R}^{>0} \times (\mathbb{R}^{>0})^r$.*

In both cases, the tangent space of $h^{-1}(0)$ is given by

$$T_{(x,\alpha,\beta)}(h^{-1}(0)) = \ker(Dh(x, \alpha, \beta)). \tag{16}$$

Proof Part a) follows from Corollary 5.14 and part b) follows from Theorem 5.12 in [16]. The formula for the tangent space follows from Proposition 5.38 in [16]. □

Remark 3 Equation (16) in the previous theorem can be used to compute tangent vectors of the regularization path in practice by computing elements of $\mathrm{pr}_x(\ker(Dh(x, \alpha, \beta)))$. Thus, it is an essential result for the construction of path-following methods.

The previous theorem is the main result in this section. It shows that the structure of $h^{-1}(0)$ (and thus the structure of $P_c^I \cap \Omega^I$ due to (13)) is related to the rank of the Jacobian Dh, given by

$$\begin{pmatrix} \alpha\nabla^2 f(x) + \sum_{j=1}^r \beta_j \nabla^2 g_{i_j}(x) & \nabla f(x) & \nabla g_{i_1}(x) & \dots & \nabla g_{i_r}(x) \\ 0 & 1 & 1 & \dots & 1 \\ (\nabla g_{i_2}(x) - \nabla g_{i_1}(x))^\top & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ (\nabla g_{i_r}(x) - \nabla g_{i_1}(x))^\top & 0 & 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{(n+r)\times(n+r+1)}$$

for $(x, \alpha, \beta) \in \mathbb{R}^n \times \mathbb{R}^{>0} \times (\mathbb{R}^{>0})^r$. Note that in Theorem 2 b), the assumption on the rank has to hold for all $(x, \alpha, \beta) \in \mathbb{R}^n \times \mathbb{R}^{>0} \times (\mathbb{R}^{>0})^r$ whereas in a), it only has to hold for all $(x, \alpha, \beta) \in h^{-1}(0)$. The following remark shows how the structure of Dh can be used to analyze its rank.

Remark 4 In the setting of Lemma 6, let $(v^x, v^\alpha, v^\beta) \in \ker(Dh(x, \alpha, \beta)) \subseteq \mathbb{R}^n \times \mathbb{R}^{>0} \times (\mathbb{R}^{>0})^r$, i.e.,

$$\begin{aligned} &\left(\alpha\nabla^2 f(x) + \sum_{j=1}^r \beta_j \nabla^2 g_{i_j}(x)\right) v^x + v^\alpha \nabla f(x) + \sum_{j=1}^r v_j^\beta \nabla g_{i_j}(x) = 0, \\ &v^\alpha + \sum_{j=1}^r v_j^\beta = 0, \\ &(\nabla g_{i_j}(x) - \nabla g_{i_1}(x))^\top v^x = 0 \quad \forall j \in \{2, \dots, r\}. \end{aligned} \tag{17}$$

Since $\{\nabla f(x), \nabla g_{i_1}(x), \ldots, \nabla g_{i_r}(x)\}$ is affinely independent by construction (cf. proof of Lemma 6), the set

$$W := \left\{ v^\alpha \nabla f(x) + \sum_{j=1}^{r} v_j^\beta \nabla g_{i_j}(x) : v^\alpha \in \mathbb{R}, v^\beta \in \mathbb{R}^r, v^\alpha + \sum_{j=1}^{r} v_j^\beta = 0 \right\}$$

is an r-dimensional linear subspace of $\mathbb{R}^n$. Similar to Lemma SM4, it is possible to show that for each element of W, the corresponding coefficients v^α and v^β are unique. If $\alpha \nabla^2 f(x) + \sum_{j=1}^{r} \beta_j \nabla^2 g_{i_j}(x)$ is regular, then the first two lines of (17) are equivalent to

$$v^x \in -\left(\alpha \nabla^2 f(x) + \sum_{j=1}^{r} \beta_j \nabla^2 g_{i_j}(x) \right)^{-1} W =: V_1,$$

where V_1 is an r-dimensional linear subspace of $\mathbb{R}^n$. In particular, v^α and v^β are uniquely determined by v^x. Furthermore, if we denote by $V^\perp$ the orthogonal complement of a subspace V, then the last line of (17) is equivalent to

$$v^x \in \operatorname{span}(\{\nabla g_{i_j}(x) - \nabla g_{i_1}(x) : j \in \{2, \ldots, r\}\})^\perp =: V_2,$$

where V_2 is an $(n - (r-1))$-dimensional subspace of $\mathbb{R}^n$ since $\{\nabla g_{i_1}(x), \ldots, \nabla g_{i_r}(x)\}$ is affinely independent. Thus, the dimension of $\ker(Dh(x, \alpha, \beta))$ is given by the dimension of the intersection $V_1 \cap V_2$. If we assume that V_1 and V_2 are generic subspaces, then we can apply a basic result from linear algebra to see that

$$\begin{aligned} \dim(\ker(Dh(x, \alpha, \beta))) &= \dim(V_1 \cap V_2) = \dim(V_1) + \dim(V_2) - \dim(V_1 + V_2) \\ &= r + (n - (r-1)) - n = 1, \end{aligned}$$

i.e., the rank of $Dh(x, \alpha, \beta)$ is full and Theorem 2(a) can be applied.

The previous remark suggests that $h^{-1}(0)$ is typically a 1-dimensional manifold such that we expect its projection $P_c^I \cap \Omega^I$ to be "1-dimensional" as well by (13). Nonetheless, we will see later that there are applications where $h^{-1}(0)$ is a higher-dimensional manifold. Furthermore, there are cases where $h^{-1}(0)$ is not a manifold at all (Note that this is not necessarily caused by the nonsmoothness of g and can also happen for smooth objective functions (cf. Example 1 in [28])). Thus, for $P_c^I \cap \Omega^I$ to have a smooth structure around a (corresponding) $x^0 \in P_c$, we have to make the following assumption:

Assumption A4 In the setting of Lemma 6, Theorem 2 can be applied, i.e.,

(a) $\operatorname{rk}(Dh(x, \alpha, \beta)) = n + r \quad \forall (x, \alpha, \beta) \in h^{-1}(0)$ or
(b) $\operatorname{rk}(Dh(x, \alpha, \beta))$ is constant $\quad \forall (x, \alpha, \beta) \in \mathbb{R}^n \times \mathbb{R}^{>0} \times (\mathbb{R}^{>0})^r$.

We conclude the discussion of the structure of $P_c^I \cap \Omega^I$ by considering the special case where f is quadratic and g is piecewise (affinely) linear. Remark A.3 in the "Appendix" shows that in this case, $P_c^I \cap \Omega^I$ is (locally) an affinely linear set around points that satisfy the assumptions of Lemma 6. This coincides with the results in [12].

3.3 The structure of P_c

After analyzing the structure of $P_c^I \cap \Omega^I$, we are now in the position to analyze the structure of the Pareto critical set P_c of (7). By (9), P_c can be written as the union of $P_c^I \cap \Omega^I$ for

all possible combinations I of selection functions. Since we already discussed the structure of the individual $P_c^I \cap \Omega^I$, the only additional nonsmooth points in P_c may arise by taking their union. More precisely, nonsmooth points may arise where the different $P_c^I \cap \Omega^I$ touch, i.e., where the set of (essentially) active selection functions changes. The following lemma yields a necessary condition for identifying such points.

Lemma 7 *Let $x^0 \in P_c$ and let $\{g_1, \ldots, g_k\}$ be a set of selection functions of g with $I^e(x^0) = \{i_1, \ldots, i_l\}$, $l \in \mathbb{N}$. If for all open neighborhoods $U \subseteq \mathbb{R}^n$ of x^0, there is some $x \in P_c \cap U$ with $I^e(x) \neq I^e(x^0)$, then there are $\alpha \geq 0$ and $\beta \in (\mathbb{R}^{\geq 0})^l$ such that $\alpha + \sum_{j=1}^{l} \beta_j = 1$,*

$$\alpha \nabla f(x^0) + \sum_{j=1}^{l} \beta_j \nabla g_{i_j}(x^0) = 0$$

and $\beta_j = 0$ for some $j \in \{1, \ldots, l\}$.

Proof See "Appendix A.4". □

A visualization of the previous lemma can be seen in Example 1: In $x^0 = (1, 0)^\top$, the sets $P_c^{\{1,2\}} \cap \Omega^{\{1,2\}}$ and $P_c^{\{1\}} \cap \Omega^{\{1\}}$ touch and there is a convex combination with a zero component (cf. (12)). In this case, this causes a kink in P_c.

Note that in general, the existence of a coefficient vector with a zero component as in Lemma 7 is not a useful criterion to find points in P_c where the active set changes. For example, by Lemma 3, if the number of essentially active selection functions in x^0 is larger than $\operatorname{affdim}(\operatorname{aff}(\{\nabla f(x^0)\} \cup \partial g(x^0)))$, then there is always a coefficient vector with a zero component. A stricter condition would be that *every* coefficient vector has a zero component, i.e., that zero is located on the relative boundary of $\operatorname{conv}(\{\nabla f(x^0)\} \cup \partial g(x^0))$ (cf. Definition SM9). By Lemma 5, this would imply that Assumption A3 cannot hold, such that $P_c^I \cap \Omega^I$ may be nonsmooth around x^0. Although the theory suggests (and we will later explicitly see this in Example 6) that this must not necessarily be the case in points where the active set changes, we believe it may be a useful criterion in practice.

Nonetheless, from a theoretical point of view, the only reliable assumption we can make to exclude points where the essentially active set changes is the following:

Assumption A5 For $x^0 \in P_c$ and a set of selection functions $\{g_1, \ldots, g_k\}$ of g, there is an open neighborhood $U \subseteq \mathbb{R}^n$ of x^0 such that

$$I^e(x) = I^e(x^0) \quad \forall x \in P_c \cap U.$$

From our considerations up to this point it follows that if $x^0 \in P_c$ is a point in which Assumptions A1 to A5 hold (for the same set of selection functions), then P_c is the projection of a smooth manifold around x^0 as in Theorem 2. An overview of all five assumptions is shown in Table 1. Unfortunately, in contrast to Assumptions A1, A2, A3 and A4, A5 is only an a posteriori condition, i.e., we already have to know P_c around x^0 to be able to check if Assumption A5 holds.

Remark 5 (a) For the development of path-following methods, it is crucial to be able to detect nonsmooth points during computation of the regularization path. If the different sets $P_c \cap \Omega^I$ are computed separately, then typically (but not necessarily), the nonsmooth points of the path are the end points of these sets (in case the path is "1-dimensional", cf. Remark 4). Thus, since path-following methods compute a pointwise approximation of the path, these end points roughly appear as points where the method fails to continue

Table 1 An overview of the five assumptions required to have a smooth structure of P_c around $x^0 \in P_c$

Let $x^0 \in P_c$.	
A1	There is an open nbd. $U \ni x^0$ and a set of sel. fct. $\{g_1, \dots, g_k\}$ of $g\|_U$ with (i) $I(x^0) = \{1, \dots, k\}$, (ii) $I^e(x) = I(x) \quad \forall x \in U$, (iii) $\operatorname{affdim}(\operatorname{aff}(\{\nabla g_i(x) : i \in \{1, \dots, k\}\}))$ $= \operatorname{affdim}(\operatorname{aff}(\{\nabla g_i(x^0) : i \in \{1, \dots, k\}\})) \; \forall x \in U$.
A2	It holds $\nabla f(x^0) \notin \operatorname{aff}(\partial g(x^0))$.
A3	Let $\{g_1, \dots, g_k\}$ be a set of selection functions of g. It exists $\{i_1, \dots, i_r\} \subseteq \{1, \dots, k\}$ and $\alpha^0 \in \mathbb{R}$, $\beta^0 \in \mathbb{R}^r$ with $\alpha^0 + \sum_{j=1}^r \beta_j^0 = 1$ such that (i) $r = \operatorname{affdim}(\operatorname{aff}(\{\nabla f(x^0)\} \cup \partial g(x^0)))$, (ii) $\{\nabla f(x^0)\} \cup \{\nabla g_i(x^0) : i \in \{i_1, \dots, i_r\}\}$ aff. ind., } (cf. Lemma 3) (iii) $\alpha^0 \nabla f(x^0) + \sum_{j=1}^r \beta_j^0 \nabla g_{i_j}(x^0) = 0$, (iv) $\alpha^0 > 0, (\beta^0)_j > 0 \; \forall j \in \{1, \dots, r\}$.
A4	Assume that A1, A2 and A3 hold and let h be defined as in Lemma 6. (a) $\operatorname{rk}(Dh(x, \alpha, \beta)) = n + r \quad \forall (x, \alpha, \beta) \in h^{-1}(0)$ **or** (b) $\operatorname{rk}(Dh(x, \alpha, \beta))$ is constant $\quad \forall (x, \alpha, \beta) \in \mathbb{R}^n \times \mathbb{R}^{>0} \times (\mathbb{R}^{>0})^r$.
A5	Let $\{g_1, \dots, g_k\}$ be a set of selection functions of g. There is an open neighborhood $U \ni x^0$ with $I^e(x) = I^e(x^0) \quad \forall x \in P_c \cap U$.

with the currently active set $I \subseteq \{1, \dots, k\}$. To find the exact nonsmooth point, one could try to find the closest point where one of the Assumptions A1 to A5 is violated. While it is not clear how this can be done numerically in our general setting, it is easier in specific applications like ℓ_1-regularization [14] (where more structure can be exploited).

(b) If Assumption A5 is violated in $x^0 \in P_c$, then there are Pareto critical points arbitrarily close to x^0 with a different (essentially) active set $I' \neq I^e(x^0)$. In practice, it may be of interest to find I'. For example, in path-following methods, I' could be used to compute the direction in which P_c continues once the nonsmoothness in x^0 was detected. To this end, let $\{g_1, \dots, g_k\}$ be the set of selection functions which are all essentially active at x^0. While it is not possible to determine I' solely from the set $\operatorname{conv}(\{\nabla f(x^0)\} \cup \partial g(x^0)) = \operatorname{conv}(\{\nabla f(x^0)\} \cup \{\nabla g_1(x^0), \dots, \nabla g_k(x^0)\})$, we can at least determine all potential candidates for I' by finding all subsets $\{i_1, \dots, i_m\} \subseteq \{1, \dots, k\}$ with

$$0 \in \operatorname{conv}(\{\nabla f(x^0)\} \cup \operatorname{conv}(\{\nabla g_{i_1}(x^0), \dots, \nabla g_{i_m}(x^0)\})).$$

(c) As the union of different $P_c^I \cap \Omega^I$ for $I \subseteq \{1, \dots, k\}$, we expect that P_c (and thus R_c by Lemma 1) is typically a "1-dimensional" set. In this case, as long as the actual regularization path R (cf. (5)) is not discrete, both R_c and R have the same "dimension". Thus, outside of kinks, we expect that R_c and R coincide locally (More precisely, we expect that for $x \in R$ there is some open set $U \subseteq \mathbb{R}^n$ with $x \in U$ such that $R \cap U = R_c \cap U$). In this way, structural result about R_c could also be applied to R in the general nonconvex case.

We conclude this section with Algorithm 1, which is an abstract path-following method for P_c based on our results (for the case where P_c is connected and "1-dimensional").

Algorithm 1 Abstract path-following method

Require: Step size $t > 0$, initial point $x^1 \in P_c$ and $I \subseteq \{1, \dots, k\}$ such that $x^1 \in P_c^I \cap \Omega^I$.
1: Initialize $i = 1$.
2: Compute the projected tangent space of $P_c^I \cap \Omega^I$ in x^i via Lemma 6 and Theorem 2 and choose a tangent vector v with $\|v\|_2 = 1$ in the current direction of continuation. (Predictor step)
3: Compute a point x^{i+1} in $P_c^I \cap \Omega^I$ close to $x^i + tv$. (Corrector step)
4: **if** the end of $P_c^I \cap \Omega^I$ is detected **then**
5: Compute the end point $\bar{x}$ of $P_c^I \cap \Omega^I$.
6: **for** all $I' \subseteq I^e(\bar{x})$, $I' \neq I$, $P_c^{I'} \cap \Omega^{I'} \neq \emptyset$ **do**
7: Restart this method with $I = I'$ and some $x^1 \in P_c^{I'} \cap \Omega^{I'}$ close to $\bar{x}$.
8: **end for**
9: **else**
10: Set $i = i + 1$ and go to step 2.
11: **end if**

Note that this algorithm is purely motivated by the structure of P_c without taking any computational regards into account. As such, to obtain a practical method for specific cases, ways to implement the steps 3 to 6 have to be further investigated.

4 Examples

In this section, we will show how our results from Sect. 3 can be used to analyze the structure of regularization paths in two common applications. These are *support vector machines* (SVMs) in data classification [2] and the *exact penalty method* in constrained optimization [5, 29].

4.1 Support vector machine

Given a *data set* $\{(x^i, y^i) : x^i \in \mathbb{R}^l, y^i \in \{-1, 1\}, i \in \{1, \dots, N\}\}$, the goal of the *support vector machine* (SVM) is to find $w \in \mathbb{R}^l$ and $b \in \mathbb{R}$ such that

$$\operatorname{sign}(w^\top x^i + b) = y^i \quad \forall i \in \{1, \dots, N\}.$$

In other words, the goal is to find a hyperplane $\{x \in \mathbb{R}^l : w^\top x + b = 0\}$ such that all x^i with $y^i = 1$ lie on one side and all x^i with $y^i = -1$ lie on the other side of the hyperplane. Since such a hyperplane may not be unique, an additional goal is to find the one where the minimal distance of the x^i to the hyperplane, also known as the *margin*, is as large as possible. One way of solving this problem is the penalization approach

$$\min_{(w,b)\in\mathbb{R}^l\times\mathbb{R}} f(w, b) + \lambda g(w, b) \tag{18}$$

for $\lambda \geq 0$ and

$$f : \mathbb{R}^l \times \mathbb{R} \to \mathbb{R}, \quad (w, b) \mapsto \frac{1}{2}\|w\|_2^2,$$
$$g : \mathbb{R}^l \times \mathbb{R} \to \mathbb{R}, \quad (w, b) \mapsto \sum_{i=1}^{N} \max\{0, 1 - y^i(w^\top x^i + b)\}.$$

Roughly speaking, minimizing g ensures that the hyperplane separates the data, while minimizing f maximizes the margin. In theory, the most favorable hyperplane would be the one with $g(w, b) = 0$ (if existent) and $f(w, b)$ as small as possible. But in practice, when working with large and noisy data sets, an imperfect separation where only few points violate the separation may be more desirable. The balance between the margin and the quality of the separation can be controlled via the parameter λ in (18), yielding a regularization path R_{SVM} as in (5) (for $n = l + 1$).

Remark 6 In the literature, the roles of f and g in problem (18) are typically reversed. The resulting problem is equivalent to our formulation with the regularization parameter $\frac{1}{\lambda}$ (except for critical points of f and g) (cf. Section 12.3.2 in [2]). Nonetheless, when the regularization path of the SVM is considered, λ in (18) is more commonly used for its parametrization.

The structure of the regularization path of the SVM was already considered in earlier works. In [11], it was shown that R_{SVM} is 1-dimensional and piecewise linear up to certain degenerate points, and a path-following method was proposed that exploits this structure. It was conjectured (without proof) that the existence of these degenerate points is related to certain properties of the data points (x^i, y^i), like having duplicates of the same point or having multiple points with the same margin. In [30], these degeneracies were analyzed further and a modified path-following method was proposed, specifically taking degenerate data sets into account. Other methods for degenerate data sets were proposed in [31–33]. In the following, we will analyze how these degeneracies relate to the nonsmooth points we characterized in our results.

Obviously, f is twice continuously differentiable and g is PC^2 with selection functions

$$\left\{(w, b) \mapsto \sum_{i \in I} 1 - y^i(w^\top x^i + b) : I \subseteq \{1, \ldots, N\}\right\}.$$

Furthermore, both f and g are convex, so R_{SVM} coincides with the critical regularization path (cf. (6)). Thus, we can apply our results from Sect. 3 to analyze the structure of R_{SVM}. Since f is quadratic and all selection functions are linear, Remark A.3 shows that the regularization path is piecewise linear up to points violating the Assumptions A1 to A5. Due to the properties of g, the Assumption A1 always holds for the SVM, as shown in Remark A.5 in the "Appendix".

In the following, we will consider the remaining Assumptions A2 to A5 in the context of the SVM and relate them to the degeneracies reported in [11]. We will do this by considering Example 1 from [30], which was specifically constructed to have a degenerate regularization path.

Example 5 Consider the data set

$$\Big\{((0.7, 0.3)^\top, 1), ((0.5, 0.5)^\top, 1), ((2, 2)^\top, -1),$$
$$((1, 3)^\top, -1), ((0.75, 0.75)^\top, 1), ((1.75, 1.75)^\top, -1)\Big\}.$$

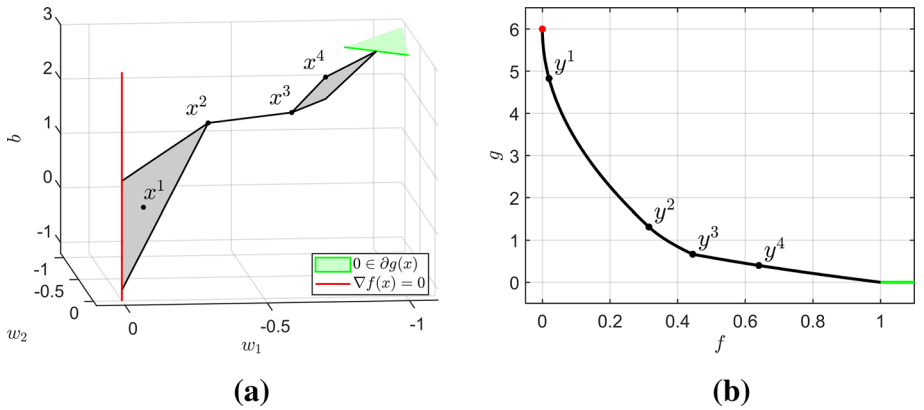

Fig. 5 **a** Regularization path of the SVM in Example 5 and the points $x^1 = \frac{1}{372}(-35, -65, 137)^\top$, $x^2 = \frac{1}{93}(-35, -65, 137)^\top$, $x^3 = \frac{1}{3}(-2, -2, 5)^\top$ and $x^4 = \frac{1}{5}(-4, -4, 11)^\top$. **b** Image of the regularization path with $y^i = (f(x^i), g(x^i))^\top$, $i \in \{1, \ldots, 4\}$, and the same coloring as in (**a**)

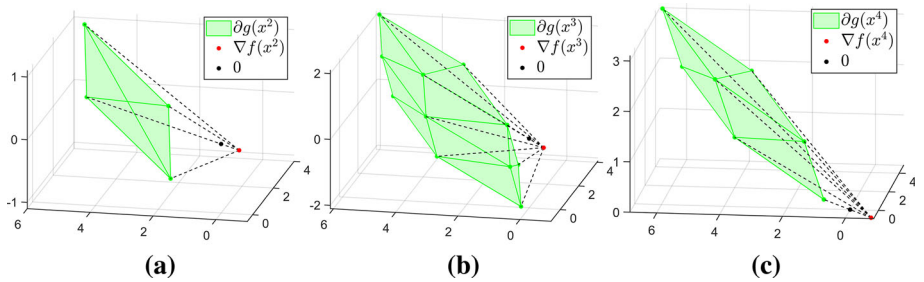

Fig. 6 Gradient of f, subdifferential of g and the (relative) boundary of the convex hull (dashed) in x^2, x^3 and x^4 in Example 5

The regularization path for this data set can be computed analytically and is shown in Fig. 5a. In the following, we will analyze the points x^1, x^2, x^3 and x^4 highlighted in Fig. 5a with respect to the Assumptions A2 to A5.

The point x^1 lies in one of the 2-dimensional parts of the regularization path and it is possible to show that g is smooth around x^1. It is easy to verify that Assumptions A2, A3 and A5 are satisfied. With regard to Assumption A4, it holds $r = \operatorname{affdim}(\operatorname{aff}(\{\nabla f(x^1)\} \cup \partial g(x^1))) = 1$ (cf. Lemma 3) and

$$Dh(x, \alpha, \beta) = \begin{pmatrix} 2\alpha & 0 & 0 & -\frac{35}{372} & \frac{14}{5} \\ 0 & 2\alpha & 0 & -\frac{65}{372} & \frac{26}{5} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

with $\operatorname{rk}(Dh(x, \alpha, \beta)) = 3$ for all $(x, \alpha, \beta) \in \mathbb{R}^n \times \mathbb{R}^{>0} \times \mathbb{R}^{>0}$. Thus, A4(b) holds which by Theorem 2 implies that the regularization path is the projection of an $n + r + 1 - m = 3 + 1 + 1 - 3 = 2$ dimensional manifold around x^1, as expected.

The point x^2 lies in a kink in the regularization path. The subdifferential of g in x^2 can be computed analytically and is shown in Fig. 6(a).

In this case, we have $\operatorname{affdim}(\operatorname{aff}(\partial g(x^2))) = 2$ and $\nabla f(x^2) \notin \operatorname{aff}(\partial g(x^2))$, so Assumption A2 holds. We see that zero lies on the relative boundary of $\operatorname{conv}(\{\nabla f(x^2) \cup \partial g(x^2)\})$ such that Assumption A3 must be violated (by Lemma 5). Furthermore, it is possible to show that the active set changes in x^2, so Assumption A5 is violated as well.

The point x^3 lies in another kink of the regularization path. The corresponding subdifferential of g is shown in Fig. 6b. As for x^2, Assumptions A3 and A5 are violated in x^3. But in contrast to x^2 we have $\operatorname{affdim}(\operatorname{aff}(\partial g(x^2))) = 3$, so $\nabla f(x^2) \in \operatorname{aff}(\partial g(x^2)) = \mathbb{R}^3$ trivially holds and Assumption A2 is violated. As discussed in Remark 2, this results in a kink in the Pareto front in the image of x^3 under the objective vector (f, g), as can be seen in Fig. 5b.

Finally, x^4 marks a corner of one of the 2-dimensional parts of the regularization path and the corresponding subdifferential is shown in Fig. 6c. As for x^3, Assumptions A2, A3 and A5 are violated in x^4. But unlike x^3, when we consider the image of x^4 in Fig. 5b, we see that there is no kink in y^4. This suggests that the KKT multiplier of f is unique even though Assumption A2 is violated. Note that this is not a contradiction to Lemma 4 b), as 0 lies on the relative boundary of $\operatorname{conv}(\{\nabla f(x^4)\} \cup \partial g(x^4))$.

4.2 Exact penalty method

Consider the constrained optimization problem

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} f(x) \\ s.t.\quad &c_i^1(x) \leq 0, \quad i \in \{1, \dots, p\}, \\ &c_j^2(x) = 0, \quad j \in \{1, \dots, q\}, \end{aligned} \tag{19}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $c_i^1 : \mathbb{R}^n \to \mathbb{R}$, $i \in \{1, \dots, p\}$, and $c_j^2 : \mathbb{R}^n \to \mathbb{R}$, $j \in \{1, \dots, q\}$, are continuously differentiable. In order to solve (19) the so-called *exact penalty method* can be used, where the idea is to solve the (nonsmooth) problem

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda g(x) \tag{20}$$

with a penalty weight $\lambda \geq 0$ and

$$g : \mathbb{R}^n \to \mathbb{R}, \quad x \mapsto \left(\sum_{i=1}^{p} \max(c_i^1(x), 0) + \sum_{j=1}^{q} |c_j^2(x)| \right).$$

It is easy to see that g is PC^1 and that a set of selection functions is given by

$$\left\{ g_{\theta,\sigma} : \mathbb{R}^n \to \mathbb{R},\ x \mapsto \sum_{i=1}^{p} \theta_i c_i^1(x) + \sum_{j=1}^{q} \sigma_i c_j^2(x) : \theta \in \{0, 1\}^p, \sigma \in \{-1, 1\}^q \right\}. \tag{21}$$

The method is based on the theoretical result that there is some $\bar{\lambda} > 0$ such that every strict local minimizer of (19) is a local minimizer of (20) for every $\lambda > \bar{\lambda}$, i.e., if λ is large enough, then the constrained problem (19) can be solved via the unconstrained problem (20) (cf. [5], Theorem 17.3). In practice, problem (20) will become ill-conditioned if λ is large compared to $\bar{\lambda}$. Thus, it is instead solved for multiple, increasing values of λ until a feasible solution is found. This results in a regularization path R as in (5). Note that all feasible points of (19) are critical points of g and the minimizer of (19) is typically the first intersection of the

regularization path with the feasible set (when starting in the minimizer of f). In particular, the existence of $\bar{\lambda}$ as above implies that the minimizer of (19) is contained in R.

In [34], R is analyzed for the case where f is quadratic (and strictly convex) and all c_i^1 and c_j^2 are affinely linear. In this case, R coincides with the critical regularization path R_c (cf. (6)). It is shown that R is piecewise linear, which coincides with our results in Remark A.3. In [13], the more general case where f and all c_i^1 are convex and all c_j^2 are affinely linear is considered. There, it still holds $R = R_c$ and it is shown that R is piecewise smooth with kinks occurring where the constraints become satisfied or violated.

Here, we want to use our theory to analyze the critical regularization path R_c in the more general setting where f, c_i^1 and c_j^2 are merely continuously differentiable. By our results in Sect. 3, we know that R_c is piecewise smooth up to points where the Assumptions A1 to A5 are violated. In Remark A.6 in the "Appendix", it is shown that if all $x \in \mathbb{R}^n$ satisfy the *linear independence constraint qualification* (LICQ), i.e., if

$$\{\nabla c_i^1(x) : c_i^1(x) = 0\} \cup \{\nabla c_j^2(x) : c_j^2(x) = 0\} \tag{22}$$

is linearly independent for all $x \in \mathbb{R}^n$, then Assumption A1 always holds and only Assumptions A2 to A5 may cause nonsmoothness in R_c. For these remaining assumptions we consider the following example, where the feasible set is given by continuously differentiable but nonconvex inequality constraints. It is inspired by problem (15) in [13].

Example 6 Consider the constrained optimization problem (19) with

$$\begin{aligned}
f(x) &= \frac{1}{2}x_1^2 + x_2^2 - x_1x_2 + \frac{1}{2}x_1 - 2x_2,\\
c_1^1(x) &= -\left(\left(x_1 - \frac{1}{2}\right)^2 + x_2^2 - 1\right),\\
c_2^1(x) &= \left(x_1 + \frac{1}{2}\right)^2 + x_2^2 - 1,\\
c_3^1(x) &= -\left(x_1^2 + \left(x_2 - \frac{1}{2}\right)^2 - 1\right).
\end{aligned}$$

The corresponding critical regularization path R_c of (20) can be computed analytically and is shown in black in Fig. 7a, consisting of two disconnected paths. The feasible set of the constrained problem coincides with the critical set of g, excluding the three isolated critical points of g. Since c_1^1 and c_3^1 are nonconvex, g is nonconvex as well, which is why R_c does not coincide with the actual regularization path R in this case. More precisely, R is merely the union of the path from the minimal point of f to x^2 and the intersection of R_c with the feasible set (cf. Fig. 7).

In the following we will analyze the kinks of R_c, which are located in x^1 to x^4 and between the minimal point of f and x^1 (cf. Fig. 7a). First of all, it is easy to see that kinks occur precisely where constraints become satisfied or violated along R_c. Due to the construction of the selection functions (cf. (21)), this causes Assumption A5 to be violated in these points.

For x^1, the gradient of f and the subdifferential of g are shown in Fig. 8a. We see that Assumption A2 holds and that Assumption A3 is violated since zero lies on the relative boundary of $\operatorname{conv}(\{\nabla f(x^1)\} \cup \partial g(x^1))$ (cf. Lemma 5). The same behavior occurs in all other kinks except for x^2. For x^2, $\nabla f(x^2)$ and $\partial g(x^2)$ are shown in Fig. 8b. In contrast to the other points, Assumption A2 is clearly violated since $\dim(\operatorname{aff}(\partial g(x^2))) = 2 = n$. As discussed in Remark 2, this causes a kink in the image of R_c, which can be seen in Fig. 7b. Moreover, zero

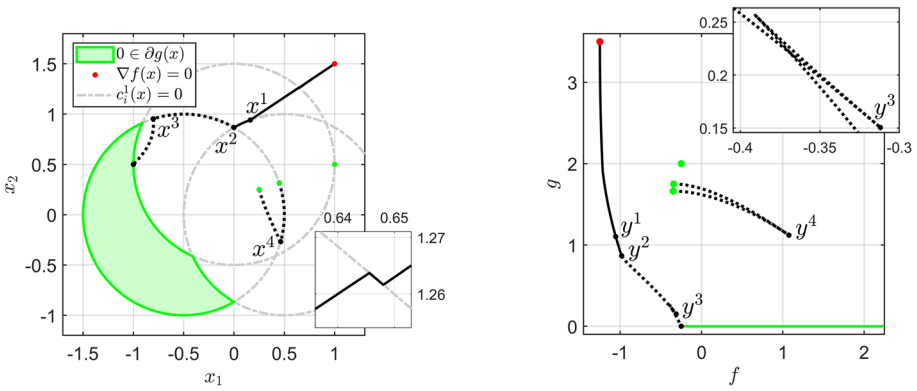

Fig. 7 **a** R (black, solid) and R_c (black) for the exact penalty method in Example 6 and the points $x^1 \approx (0.1614, 0.9409)^\top$, $x^2 = (0, \frac{\sqrt{3}}{2})^\top$, $x^3 \approx (-0.8027, 0.9531)^\top$, $x^4 \approx (0.4631, -0.2691)^\top$ with a zoom of the intersection of $c_3^1(x) = 0$ and R_c. **b** Image of R_c with $y^i = (f(x^i), g(x^i)), i \in \{1, \ldots, 4\}$, and the same coloring as in (**a**). Furthermore, a zoom of the image around y^3

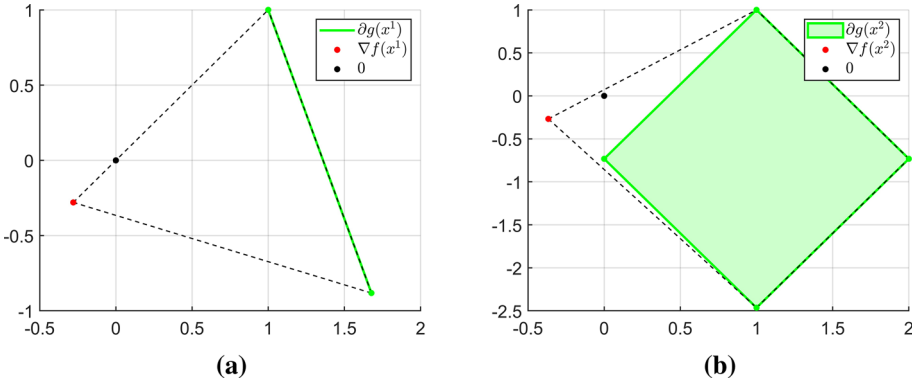

Fig. 8 Gradient of f, subdifferential of g and the corresponding (relative) boundary of the convex hull (dashed) in x^1 and x^2 of Example 6

lies in the relative interior of $\text{conv}(\{\nabla f(x^2)\} \cup \partial g(x^2))$ and it is easy to see that Assumption A3 holds.

In addition to the features described so far, the image of R_c possesses so-called *turning points*. If we treat the image of R_c as an actual (continuous) path, then these are points where the direction of the path abruptly turns around. For example, this can be observed in y^3 and y^4 in Fig. 7b. These points were already discussed in [14] and in Example 3.4 therein, it was highlighted that they are not necessarily caused by any nonsmoothess of the objectives. Since we are mainly interested in the structure of R_c in this article, we will leave their analysis for future work.

Note that all kinks in the previous examples were points where constraints become satisfied or violated, which suggests that the structural results from [13] also hold in our more general nonconvex case, at least for the critical regularization path R_c. Furthermore, R_c is still connecting the minimum of f to the solution of the constrained problem (19) (which is the

intersection of R_c with the feasible set). Thus, it might be possible to apply a path-following method similar to the one in [13] to nonconvex problems as well.

5 Conclusion

In this article, we have presented results about the structure of regularization paths for piecewise differentiable regularization terms. We did this by first showing that the critical regularization path is related to the Pareto critical set P_c of the multiobjective optimization problem which contains the objective function f and the regularization term g. Afterwards, we analyzed P_c by reformulating it as a union of the intersection of certain sets, which allowed us to apply differential geometry to obtain structural results. During this derivation, we identified five assumptions (A1 to A5) which (when combined) are sufficient for P_c to have a smooth structure locally around a given $x^0 \in P_c$. In turn, nonsmooth features of P_c (like "kinks") can be classified depending on which of these five assumptions is violated. We demonstrated this by analyzing the regularization paths for the support-vector machine and the exact penalty method.

Based on our results in this article, there are multiple possible directions for future work:

- We believe that most of our theoretical results would still hold (with only minor adjustments) if we would assume f to be merely piecewise differentiable as well (In this case, the regularization function $f + \lambda g$ would still be piecewise differentiable).
- Although the MOP (7) considered in this article has only two objectives, multiobjective optimization can handle any number of objectives. In particular, (7) could be formulated for arbitrarily many regularization terms. We believe that results similar to ours (with a higher-dimensional regularization path) could be obtained for this case. This would allow regularization methods such as the *elastic net* [35] to be incorporated into our framework.
- While we were focused on regularization in this article, our results can also be used in the context of general multiobjective optimization to construct path-following methods for the solution of nonsmooth MOPs, extending [12–14, 26].
- Although we provided the main ingredients for the construction of path-following methods, i.e., a way to compute the tangent space in smooth areas and a characterization of nonsmooth points, their development and actual implementation is still non-trivial. For example, other important ingredients are the computation of new points on R_c after taking a step along the tangent direction (also known as a *corrector*), the detection of kinks in the path and the computation of the correct tangent direction after a kink was found. Treating these problems in our general framework could greatly simplify the development of new path-following methods.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10898-022-01223-2.

Acknowledgements This research has been funded by the European Union and the German Federal State of North Rhine-Westphalia within the EFRE.NRW Project "SET CPS", and by the DFG Priority Programme 1962 "Non-smooth and Complementarity-based Distributed Parameter Systems".

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability Data sharing not applicable—no new data generated

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Appendix A

A.1 Proof of Lemma 4

Let $\{g_1, \ldots, g_k\}$ be a set of selection functions of g and let $I^e(x^0) = \{i_1, \ldots, i_l\}$.

(a) By assumption, for $s \in \{1, 2\}$, there have to be $\alpha_1^s > 0$ and $\beta^s \in (\mathbb{R}^{\geq 0})^l$ such that $\alpha_1^s + \sum_{j=1}^l \beta_j^s = 1$,

$$\alpha_1^s \nabla f(x^0) + \sum_{j=1}^{l} \beta_j^s \nabla g_{i_j}(x^0) = 0 \tag{A1}$$

and $\alpha_1^1 \neq \alpha_1^2$. This implies

$$\alpha_1^1 \nabla f(x^0) + \sum_{j=1}^{l} \beta_j^1 \nabla g_{i_j}(x^0) = \alpha_1^2 \nabla f(x^0) + \sum_{j=1}^{l} \beta_j^2 \nabla g_{i_j}(x^0)$$

$$\Leftrightarrow \nabla f(x^0) = \frac{1}{\alpha_1^1 - \alpha_1^2} \sum_{j=1}^{l} (\beta_j^2 - \beta_j^1) \nabla g_{i_j}(x^0) = \sum_{j=1}^{l} \frac{\beta_j^2 - \beta_j^1}{\alpha_1^1 - \alpha_1^2} \nabla g_{i_j}(x^0)$$

with

$$\sum_{j=1}^{l} \frac{\beta_j^2 - \beta_j^1}{\alpha_1^1 - \alpha_1^2} = \frac{1 - \alpha_1^2 - (1 - \alpha_1^1)}{\alpha_1^1 - \alpha_1^2} = 1,$$

showing that $\nabla f(x^0) \in \operatorname{aff}(\partial g(x^0))$.

(b) Since $\nabla f(x^0) \in \operatorname{aff}(\partial g(x^0))$ there has to be some $\beta' \in \mathbb{R}^l$ with $\sum_{j=1}^l \beta'_j = 1$ and

$$\nabla f(x^0) = \sum_{j=1}^{l} \beta'_j \nabla g_{i_j}(x^0). \tag{A2}$$

Furthermore, by Lemma SM5, zero being contained in $\mathrm{ri}(\mathrm{conv}(\{\nabla f(x^0)\} \cup \partial g(x^0)))$ is equivalent to the existence of $\alpha_1 > 0$ and $\beta \in (\mathbb{R}^{>0})^l$ with $\alpha_1 + \sum_{j=1}^{l} \beta_j = 1$ and

$$\alpha_1 \nabla f(x^0) + \sum_{j=1}^{l} \beta_j \nabla g_{i_j}(x^0) = 0. \tag{A3}$$

Combination of (A2) and (A3) yields

$$(\alpha_1 - \lambda)\nabla f(x^0) + \sum_{j=1}^{l} (\beta_j + \lambda\beta'_j)\nabla g_{i_j}(x^0) = 0 \quad \forall \lambda \in \mathbb{R}$$

and

$$(\alpha_1 - \lambda) + \sum_{j=1}^{l} (\beta_j + \lambda\beta'_j) = \alpha_1 + \sum_{j=1}^{l} \beta_j + \lambda \left(-1 + \sum_{j=1}^{l} \beta'_j \right) = 1 \quad \forall \lambda \in \mathbb{R}.$$

Since $\alpha_1 > 0$ and $\beta \in (\mathbb{R}^{>0})^l$, there has to be some $\lambda \neq 0$ such that $\alpha_1 - \lambda > 0$ and $\beta + \lambda\beta' \in (\mathbb{R}^{>0})^l$. In particular, $\alpha_1 - \lambda \neq \alpha_1$ is another KKT multiplier corresponding to f in x^0, completing the proof.

A.2 Proof of Lemma 5

Let $\{g_1, \ldots, g_k\}$ be a set of selection functions that satisfies A3. Let $L := \{1, \ldots, k\} \setminus \{i_1, \ldots, i_r\}$. Since $\{\nabla f(x^0)\} \cup \{\nabla g_i(x^0) : i \in \{i_1, \ldots, i_r\}\}$ is an affine basis of $\mathrm{aff}(\{\nabla f(x^0)\} \cup \partial g(x^0))$, there are coefficients $\theta^l \in \mathbb{R}$ and $\nu^l \in \mathbb{R}^r$ for every $l \in L$ with $\theta^l + \sum_{j=1}^{r} \nu_j^l = 1$ and

$$\nabla g_l(x^0) = \theta^l \nabla f(x^0) + \sum_{j=1}^{r} \nu_j^l \nabla g_{i_j}(x^0).$$

Let $\bar{\theta} := -\sum_{l \in L} \theta^l$ and $\bar{\nu}_j := -\sum_{l \in L} \nu_j^l$ for $j \in \{1, \ldots, r\}$. Then

$$\begin{aligned} 0 &= \sum_{l \in L} \left(\nabla g_l(x^0) - \theta^l \nabla f(x^0) - \sum_{j=1}^{r} \nu_j^l \nabla g_{i_j}(x^0) \right) \\ &= \bar{\theta} \nabla f(x^0) + \sum_{j=1}^{r} \bar{\nu}_j \nabla g_{i_j}(x^0) + \sum_{l \in L} \nabla g_l(x^0) \end{aligned}$$

and $\bar{\theta} + \sum_{j=1}^{r} \bar{\nu}_j + \sum_{l \in L} 1 = 0$. Let $\alpha^0 > 0$ and $\beta^0 \in (\mathbb{R}^{>0})^r$ as in A3. Then

$$\begin{aligned} 0 &= \alpha^0 \nabla f(x^0) + \sum_{j=1}^{r} \beta_j^0 \nabla g_{i_j}(x^0) \\ &= \alpha^0 \nabla f(x^0) + \sum_{j=1}^{r} \beta_j^0 \nabla g_{i_j}(x^0) + \lambda \left(\bar{\theta} \nabla f(x^0) + \sum_{j=1}^{r} \bar{\nu}_j \nabla g_{i_j}(x^0) + \sum_{l \in L} \nabla g_l(x^0) \right) \\ &= (\alpha^0 + \lambda\bar{\theta}) \nabla f(x^0) + \sum_{j=1}^{r} (\beta_j^0 + \lambda \bar{\nu}_j) \nabla g_{i_j}(x^0) + \sum_{l \in L} \lambda \nabla g_l(x^0) \end{aligned} \tag{A4}$$

for all $\lambda \in \mathbb{R}$. By construction, there is some $\lambda > 0$ such that (A4) is a vanishing convex combination with strictly positive coefficients. Applying Lemma SM5 completes the proof.

A.3 Remark regarding Sect. 3.2

Let

$$f : \mathbb{R}^n \to \mathbb{R}, \quad x \mapsto \frac{1}{2} x^\top A x + b^\top x + c$$

for $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Furthermore, assume that there is a set of selection functions $\{g_1, \ldots, g_k\}$ of g consisting of affinely linear functions, i.e.,

$$g_i : \mathbb{R}^n \to \mathbb{R}, \quad x \mapsto d_i^\top x + e_i$$

for $d_i \in \mathbb{R}^n$, $e_i \in \mathbb{R}$, $i \in \{1, \ldots, k\}$. Let $x^0 \in P_c$ and assume that Lemma 6 is applicable, yielding an index set $\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, k\}$, an open neighborhood $U' \subseteq \mathbb{R}^n$ of x^0 and coefficients $\alpha^0 \in \mathbb{R}^{>0}$ and $\beta^0 \in (\mathbb{R}^{>0})^r$ such that $h(x^0, \alpha^0, \beta^0) = 0$. In this case, the map h reduces to

$$h(x, \alpha, \beta) = \begin{pmatrix} \alpha(Ax + b) + \sum_{j=1}^r \beta_j d_{i_j} \\ \alpha + \sum_{j=1}^r \beta_j - 1 \\ ((d_{i_j}^\top - d_{i_1})^\top x + e_{i_j} - e_{i_1})_{j \in \{2, \ldots, r\}} \end{pmatrix}.$$

We will show that

$$\mathrm{pr}_x(h^{-1}(0)) \cap U' = (x^0 + \mathrm{pr}_x(\ker(Dh(x^0, \alpha^0, \beta^0)))) \cap U', \tag{A5}$$

which by Lemma 6 implies that $P_c^{I^e(x^0)} \cap \Omega^{I^e(x^0)} \cap U'$ is an affinely linear set with dimension $\dim(\mathrm{pr}_x(\ker(Dh(x^0, \alpha^0, \beta^0))))$.

To this end, let $(v^x, v^\alpha, v^\beta) \in \ker(Dh(x^0, \alpha^0, \beta^0))$. Since $\alpha^0 > 0$, there is some $\varepsilon > 0$ such that $\alpha^0 - t v^\alpha > 0$ for all $t \in [0, \varepsilon)$. Define

$$s : [0, \varepsilon) \to \mathbb{R}, \quad t \mapsto \frac{t \alpha^0}{\alpha^0 - t v^\alpha}.$$

Since $\alpha^0 > 0$ and $\beta^0 \in (\mathbb{R}^{>0})^r$, there is some $\varepsilon' \in (0, \varepsilon)$ such that

$$\begin{aligned} &\alpha^0 + s(t) v^\alpha > 0, \\ &\beta_j^0 + s(t) v_j^\beta > 0 \quad \forall j \in \{1, \ldots, r\} \end{aligned}$$

for all $t \in [0, \varepsilon')$. Furthermore, since U' is an open neighborhood of x^0, there is some $\varepsilon'' > 0$ such that $x^0 + t v^x \in U'$ for all $t \in [0, \varepsilon'')$. Finally, a simple calculation shows that

$$h(x^0 + t v^x, \alpha^0 + s(t) v^\alpha, \beta^0 + s(t) v^\beta) = 0 \quad \forall t \in [0, \varepsilon'').$$

Thus, "$\supseteq$" holds in (A5).
In turn, let $(x^1, \alpha^1, \beta^1) \in U' \times \mathbb{R}^{>0} \times (\mathbb{R}^{>0})^r$ with $h(x^1, \alpha^1, \beta^1) = 0$. Let $s := \frac{\alpha^0}{\alpha^1}$. It is easy to show that

$$(x^1 - x^0, s(\alpha^1 - \alpha^0), s(\beta^1 - \beta^0)) \in \ker(Dh(x^0, \alpha^0, \beta^0)),$$

implying that "$\subseteq$" holds in (A5).

A.4 Proof of Lemma 7

By assumption there is a sequence $(x^s)_{s\in\mathbb{N}} \in P_c$ with $\lim_{s\to\infty} x^s = x^0$ and $I^e(x^s) \neq I^e(x^0)$ for all $s \in \mathbb{N}$. Assume w.l.o.g. that $I^e(x^s)$ is constant for all $s \in \mathbb{N}$. Due to the definition of the essentially active set, we can assume w.l.o.g. that $I^e(x^s) = \{i_1, \dots, i_m\} \subseteq I^e(x^0)$ for some $m < l$. Since $x^s \in P_c$ for all $s \in \mathbb{N}$, there are sequences $(\alpha^s)_{s\in\mathbb{N}} \in \mathbb{R}^{\geq 0}$, $(\beta^s)_{s\in\mathbb{N}} \in (\mathbb{R}^{\geq 0})^m$ with $\alpha^s + \sum_{j=1}^m \beta_j^s = 1$ and

$$\alpha^s \nabla f(x^s) + \sum_{j=1}^m \beta_j^s \nabla g_{i_j}(x^s) = 0$$

for all $s \in \mathbb{N}$. Since $(\alpha^s)_{s\in\mathbb{N}}$ and $(\beta^s)_{s\in\mathbb{N}}$ are bounded, we can assume w.l.o.g. that there are $\alpha \in \mathbb{R}^{\geq 0}$ and $\bar{\beta} \in (\mathbb{R}^{\geq 0})^m$ with $\lim_{s\to\infty} \alpha^s = \alpha$ and $\lim_{s\to\infty} \beta^s = \bar{\beta}$. In particular, $\alpha + \sum_{j=1}^l \bar{\beta}_j = 1$. By continuity of ∇f and ∇g_i, $i \in \{1, \dots, k\}$, we have

$$\alpha \nabla f(x^0) + \sum_{j=1}^m \bar{\beta}_j \nabla g_{i_j}(x^0) = 0.$$

The proof follows by setting $\beta = (\bar{\beta}_{i_1}, \dots, \bar{\beta}_{i_m}, 0, \dots, 0)^\top \in (\mathbb{R}^{\geq 0})^l$.

A.5 Remark regarding Sect. 4.1

Let $g : \mathbb{R}^n \to \mathbb{R}$ be any piecewise linear and convex function. Let $x^0 \in \mathbb{R}^n$. By Lemma SM1, there is an open neighborhood $U \subseteq \mathbb{R}^n$ of x^0 and a set of (affinely linear) selection functions $\{g_1, \dots, g_k\}$ of $g|_U$ which are all essentially active in x^0. In particular, $I(x^0) = \{1, \dots, k\}$, so A1(i) holds.
To see that A1(ii) holds, let $z \in U$ and $j \in I(z)$. Since all selection functions are essentially active in x^0, we have

$$x^0 \in \mathrm{cl}(\mathrm{int}(\{y \in U : g(y) = g_j(y)\})),$$

so $V := \mathrm{int}(\{y \in U : g(y) = g_j(y)\}) \neq \emptyset$. Let $y \in V$. Since g is convex and g_j is affinely linear, we have

$$\begin{aligned} g((1-\lambda)y + \lambda z) &\leq (1-\lambda)g(y) + \lambda g(z) = (1-\lambda)g_j(y) + \lambda g_j(z) \\ &= g_j((1-\lambda)y + \lambda z) \quad \forall \lambda \in [0,1]. \end{aligned} \tag{A6}$$

Assume that we have inequality in (A6), i.e., assume that there is some $\bar{\lambda} \in [0,1]$ with $g(\bar{x}) < g_j(\bar{x})$ for $\bar{x} := (1-\bar{\lambda})y + \bar{\lambda} z$. Then

$$\begin{aligned} g((1-\lambda)y + \lambda\bar{x}) &\leq (1-\lambda)g(y) + \lambda g(\bar{x}) < (1-\lambda)g(y) + \lambda g_j(\bar{x}) \\ &= g_j((1-\lambda)y + \lambda\bar{x}) \quad \forall \lambda \in (0,1]. \end{aligned}$$

This is a contradiction to the openness of V, so we must have equality in (A6). This implies

$$j \in I((1-\lambda)y + \lambda z) \quad \forall \lambda \in [0,1].$$

As this holds for arbitrary $y \in V$, we have

$$j \in I(x) \quad \forall x \in \mathrm{conv}(V \cup \{z\}).$$

Since V is open in $\mathbb{R}^n$, it is possible to show that

$$z \in \mathrm{cl}(\mathrm{int}(\mathrm{conv}(V \cup \{z\}))) \subseteq \mathrm{cl}(\mathrm{int}(\{y \in U : g(y) = g_j(y)\})),$$

showing that $j \in I^e(z)$.
Since ∇g_i is constant for all $i \in \{1, \ldots, k\}$, it is easy to see that A1(iii) holds as well.

A.6 Remark regarding Sect. 4.2

We begin by deriving an explicit expression for the active set. To this end, let $x \in \mathbb{R}^n$ and assume w.l.o.g. that there are $\bar{p} \in \{1, \ldots, p\}$, $\bar{q} \in \{1, \ldots, q\}$ such that

$$\begin{aligned} c_i^1(x) = 0, \quad \forall i \in \{1, \ldots, \bar{p}\}, \qquad c_i^1(x) \neq 0, \quad \forall i \in \{\bar{p}+1, \ldots, p\}, \\ c_j^2(x) = 0, \quad \forall j \in \{1, \ldots, \bar{q}\}, \qquad c_j^2(x) \neq 0, \quad \forall j \in \{\bar{q}+1, \ldots, q\}. \end{aligned} \tag{A7}$$

For $i \in \{\bar{p}+1, \ldots, p\}$ and $j \in \{\bar{q}+1, \ldots, q\}$ define

$$\begin{aligned} \hat{\theta}_i &:= \begin{cases} 1, & \text{if } c_i^1(x) > 0 \\ 0, & \text{if } c_i^1(x) < 0 \end{cases}, \\ \hat{\sigma}_j &:= \mathrm{sign}(c_j^2(x)), \end{aligned} \tag{A8}$$

and

$$\bar{c} : \mathbb{R}^n \to \mathbb{R}, \quad x \mapsto \sum_{i=\bar{p}+1}^{p} \hat{\theta}_i c_i^1(x) + \sum_{j=\bar{q}+1}^{q} \hat{\sigma}_j c_j^2(x).$$

Then by construction,

$$\begin{aligned} g(x) &= \sum_{i=1}^{p} \max\{c_i^1(x), 0\} + \sum_{j=1}^{q} |c_j^2(x)| = \sum_{i=\bar{p}+1}^{p} \max\{c_i^1(x), 0\} + \sum_{j=\bar{q}+1}^{q} |c_j^2(x)| \\ &= \sum_{i=\bar{p}+1}^{p} \hat{\theta}_i c_i^1(x) + \sum_{j=\bar{q}+1}^{q} \hat{\sigma}_j c_j^2(x) = \bar{c}(x) \\ &= \bar{c}(x) + \sum_{i=1}^{\bar{p}} \bar{\theta}_i c_i^1(x) + \sum_{j=1}^{\bar{q}} \bar{\sigma}_j c_j^2(x) \\ &= g_{(\bar{\theta}, \hat{\theta}), (\bar{\sigma}, \hat{\sigma})}(x) \end{aligned}$$

for all $\bar{\theta} \in \{0, 1\}^p$, $\bar{\sigma} \in \{-1, 1\}^q$. Thus

$$\bar{I} := \left\{ ((\bar{\theta}, \hat{\theta})^\top, (\bar{\sigma}, \hat{\sigma})^\top) : \bar{\theta} \in \{0, 1\}^{\bar{p}}, \bar{\sigma} \in \{-1, 1\}^{\bar{q}} \right\} \subseteq I(x).$$

To show that "$\supseteq$" holds, let $(\tilde{\theta}, \tilde{\sigma}) \in I(x)$. Then

$$\hat{\theta}_i - \tilde{\theta}_i = \begin{cases} -1, & \text{if } \hat{\theta}_i \neq \tilde{\theta}_i,\ \hat{\theta}_i = 0 \\ 1, & \text{if } \hat{\theta}_i \neq \tilde{\theta}_i,\ \hat{\theta}_i = 1 \\ 0, & \text{otherwise} \end{cases}, \quad \hat{\sigma}_j - \tilde{\sigma}_j = \begin{cases} -2, & \text{if } \hat{\sigma}_j \neq \tilde{\sigma}_j,\ \hat{\sigma}_i = -1 \\ 2, & \text{if } \hat{\sigma}_j \neq \tilde{\sigma}_j,\ \hat{\sigma}_i = 1 \\ 0, & \text{otherwise} \end{cases}$$

for all $i \in \{\bar{p}+1, \ldots, p\}$, $j \in \{\bar{q}+1, \ldots, q\}$. Combined with (A8), this implies

$$\begin{aligned} 0 = g(x) - g_{\tilde{\theta},\tilde{\sigma}}(x) &= \sum_{i=1}^{p} \max\{c_i^1(x), 0\} + \sum_{j=1}^{q} |c_j^2(x)| - \sum_{i=1}^{p} \tilde{\theta}_i c_i^1(x) - \sum_{j=1}^{q} \tilde{\sigma}_j c_j^2(x) \\ &= \sum_{i=\bar{p}+1}^{p} (\hat{\theta}_i - \tilde{\theta}_i) c_i^1(x) + \sum_{j=\bar{q}+1}^{q} (\hat{\sigma}_j - \tilde{\sigma}_j) c_j^2(x) \\ &= \sum_{\substack{i=\bar{p}+1 \\ \hat{\theta}_i \neq \tilde{\theta}_i}}^{p} |c_i^1(x)| + \sum_{\substack{j=\bar{q}+1 \\ \hat{\sigma}_j \neq \tilde{\sigma}_j}}^{q} 2|c_j^2(x)|, \end{aligned}$$

so both sums must be empty, i.e., $\hat{\theta}_i = \tilde{\theta}_i$ for all $i \in \{\bar{p}+1, \ldots, p\}$ and $\hat{\sigma}_j = \tilde{\sigma}_j$ for all $i \in \{\bar{q}+1, \ldots, q\}$. In particular $(\tilde{\theta}, \tilde{\sigma}) \in \bar{I}$, so $\bar{I} = I(x)$ for all $x \in \mathbb{R}^n$.

In the following, we will show that all active selection functions are essentially active. To this end, let $(\theta, \sigma) \in I(x) = \bar{I}$. Define

$$\begin{aligned} v^i &:= \begin{cases} \nabla c_i^1(x), & \text{if } \theta_i = 0 \\ -\nabla c_i^1(x), & \text{if } \theta_i = 1 \end{cases} \quad \forall i \in \{1, \ldots, \bar{p}\}, \\ w^j &:= -\sigma_j \nabla c_j^2(x) \quad \forall j \in \{1, \ldots, \bar{q}\}, \\ C &:= \operatorname{conv}(\{v^i : i \in \{1, \ldots, \bar{p}\}\} \cup \{w^j : j \in \{1, \ldots, \bar{q}\}\}). \end{aligned}$$

The LICQ (cf. (22)) implies that $0 \notin C$. With a basic result from convex analysis (cf. Lemma in [36]), it follows that there is some $d \in \mathbb{R}^n \setminus \{0\}$ with

$$\begin{aligned} 0 > \langle v^i, d\rangle &= \begin{cases} \langle \nabla c_i^1(x), d\rangle, & \text{if } \theta_i = 0 \\ -\langle \nabla c_i^1(x), d\rangle, & \text{if } \theta_i = 1 \end{cases} \quad \forall i \in \{1, \ldots, \bar{p}\}, \\ 0 > \langle w^j, d\rangle &= -\langle \sigma_j \nabla c_j^2(x), d\rangle \quad \forall j \in \{1, \ldots, \bar{q}\}. \end{aligned}$$

The continuity of the constraint functions implies that there is some $T > 0$ such that

$$\begin{aligned} \operatorname{sign}(c_i^1(x + td)) &= \begin{cases} -1, & \text{if } \theta_i = 0 \\ 1, & \text{if } \theta_i = 1 \end{cases} \quad \forall i \in \{1, \ldots, p\}, \\ \operatorname{sign}(c_j^2(x + td)) &= \sigma_j \quad \forall j \in \{1, \ldots, q\}, \end{aligned}$$

for all $t \in (0, T)$. Note that in particular, for all points $x + td$ with $t \in (0, T)$, there is a neighborhood of $x + td$ on which g is smooth with $g = g_{\theta,\sigma}$. This shows that $(\theta, \sigma) \in I^e(x)$.

Let $x^0 \in \mathbb{R}^n$. From our discussion up to this point it follows that A1(i) and (ii) hold for an appropriate open neighborhood U of x^0. To show that A1(iii) holds, let (θ', σ') be any element of $\bar{I} = I(x^0)$ (with $\bar{p}$ and $\bar{q}$ as in (A7)) and $z \in U$. Clearly,

$$\begin{aligned} &\operatorname{span}(\{\nabla g_{\theta,\sigma}(z) - \nabla g_{\theta',\sigma'}(z) : (\theta, \sigma) \in \bar{I}\}) \\ &\quad \subseteq \operatorname{span}(\{\nabla c_i^1(z) : i \in \{1, \ldots, \bar{p}\}\} \cup \{\nabla c_j^2(z) : j \in \{1, \ldots, \bar{q}\}\}). \end{aligned} \tag{A9}$$

We will show that we actually have equality in (A9), which implies that A1(iii) holds by the LICQ (cf. (22)). To this end, let $i' \in \{1, \ldots, \bar{p}\}$. Define

$$\tilde{\theta}_i := \begin{cases} \theta'_i, & \text{if } i \neq i' \\ 1, & \text{if } i = i', \theta'_i = 0 \\ 0, & \text{if } i = i', \theta'_i = 1 \end{cases} \quad \forall i \in \{1, \ldots, \bar{p}\}.$$

Then $(\tilde{\theta}, \sigma') \in \bar{I}$, so $g_{\tilde{\theta},\sigma'}(z) - g_{\theta',\sigma'}(z) = \pm \nabla c^1_{i'}(z)$ and $\nabla c^1_{i'}(z)$ is contained in the left-hand side of (A9). Analogously, it is possible to show that $\nabla c^2_j(z)$ is contained in the left-hand side of (A9) for all $j \in \{1, \ldots, \bar{q}\}$, such that equality holds.

References

1. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodol.) **58**(1), 267–288 (1996)
2. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, Berlin (2009). https://doi.org/10.1007/978-0-387-84858-7
3. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin (2006)
4. Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imaging Vis. **20**, 89–97 (2004). https://doi.org/10.1023/b:jmiv.0000011325.36760.1e
5. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer Series in Operations Research and Financial Engineering, 2nd edn. Springer (2006)
6. Bagirov, A., Karmitsa, N., Mäkelä, M.M.: Introduction to Nonsmooth Optimization. Springer (2014). https://doi.org/10.1007/978-3-319-08114-4
7. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016). http://www.deeplearningbook.org
8. Branke, J., Deb, K., Dierolf, H., Osswald, M.: In: Lecture Notes in Computer Science, pp. 722–731. Springer, Berlin (2004). https://doi.org/10.1007/978-3-540-30217-9_73
9. Osborne, M., Presnell, B., Turlach, B.A.: A new approach to variable selection in least squares problems. IMA J. Numer. Anal. **20**(3), 389–403 (2000). https://doi.org/10.1093/imanum/20.3.389
10. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. (2004). https://doi.org/10.1214/009053604000000067
11. Hastie, T., Rosset, S., Tibshirani, R., Zhu, J.: The entire regularization path for the support vector machine. J. Mach. Learn. Res. **5**, 1391–1415 (2004)
12. Rosset, S., Zhu, J.: Piecewise linear regularized solution paths. Ann. Stat. (2007). https://doi.org/10.1214/009053606000001370
13. Zhou, H., Lange, K.: Path following in the exact penalty method of convex programming. Comput. Optim. Appl. **61**(3), 609–634 (2015). https://doi.org/10.1007/s10589-015-9732-x
14. Bieker, K., Gebken, B., Peitz, S.: On the treatment of optimization problems with L1 penalty terms via multiobjective continuation. IEEE Trans. Pattern Anal. Mach. Intell. (2021). https://doi.org/10.1109/TPAMI.2021.3114962
15. Scholtes, S.: Introduction to Piecewise Differentiable Equations. Springer Briefs in Optimization. Springer, New York (2012). https://doi.org/10.1007/978-1-4614-4340-7
16. Lee, J.: Introduction to Smooth Manifolds, 2nd edn. Springer (2012). https://doi.org/10.1007/978-1-4419-9982-5
17. Clarke, F.H.: Optimization and Nonsmooth Analysis. Society for Industrial and Applied Mathematics (1990). https://doi.org/10.1137/1.9781611971309
18. Miettinen, K.: Nonlinear Multiobjective Optimization. Springer US (1998). https://doi.org/10.1007/978-1-4615-5563-6
19. Ehrgott, M.: Multicriteria Optimization. Springer-Verlag (2005). https://doi.org/10.1007/3-540-27659-9
20. Mäkelä, M.M., Eronen, V.P., Karmitsa, N.: 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
21. Rockafellar, R.T.: Convex Analysis. Princeton University Press (1970). https://doi.org/10.1515/9781400873173

22. Gallier, J.: Geometric Methods and Applications. Springer, New York (2011). https://doi.org/10.1007/978-1-4419-9961-0
23. Jungnickel, D.: Optimierungsmethoden. Springer, Berlin (2015). https://doi.org/10.1007/978-3-642-54821-5
24. Park, M.Y., Hastie, T.: L1-regularization path algorithm for generalized linear models. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **69**(4), 659–677 (2007). https://doi.org/10.1111/j.1467-9868.2007.00607.x
25. Mairal, J., Yu, B.: In: Proceedings of the 29th International Conference on International Conference on Machine Learning, ICML'12, pp. 1835–1842. Omnipress, Madison (2012)
26. Hillermeier, C.: Nonlinear Multiobjective Optimization. Birkhäuser, Basel (2001). https://doi.org/10.1007/978-3-0348-8280-4
27. Evans, L.C., Gariepy, R.F.: Measure Theory and Fine Properties of Functions, Revised Edition. Chapman and Hall/CRC (2015). https://doi.org/10.1201/b18333
28. Gebken, B., Peitz, S., Dellnitz, M.: On the hierarchical structure of pareto critical sets. J. Glob. Optim. **73**(4), 891–913 (2019). https://doi.org/10.1007/s10898-019-00737-6
29. Pillo, G.D., Grippo, L.: Exact penalty functions in constrained optimization. SIAM J. Control. Optim. **27**(6), 1333–1360 (1989). https://doi.org/10.1137/0327068
30. Ong, C.J., Shao, S., Yang, J.: An improved algorithm for the solution of the regularization path of support vector machine. IEEE Trans. Neural Netw. **21**(3), 451–462 (2010). https://doi.org/10.1109/tnn.2009.2039000
31. Dai, J., Chang, C., Mai, F., Zhao, D., Xu, W.: On the SVMpath singularity. IEEE Trans. Neural Netw. Learn. Syst. **24**(11), 1736–1748 (2013). https://doi.org/10.1109/tnnls.2013.2262180
32. Sentelle, C.G., Anagnostopoulos, G.C., Georgiopoulos, M.: A simple method for solving the SVM regularization path for semidefinite kernels. IEEE Trans. Neural Netw. Learn. Syst. **27**(4), 709–722 (2016). https://doi.org/10.1109/tnnls.2015.2427333
33. Wang, B., Zhou, L., Cao, Z., Dai, J.: Ridge-adding approach for SVMpath singularities. IEEE Access **7**, 47728–47736 (2019). https://doi.org/10.1109/access.2019.2909297
34. Zhou, H., Lange, K.: A path algorithm for constrained estimation. J. Comput. Graph. Stat. **22**(2), 261–283 (2013)
35. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **67**(2), 301–320 (2005). https://doi.org/10.1111/j.1467-9868.2005.00503.x
36. Cheney, W., Goldstein, A.A.: Proximity maps for convex sets. Proc. Am. Math. Soc. **10**(3), 448–448 (1959). https://doi.org/10.1090/s0002-9939-1959-0105008-8
37. Ulbrich, M.: 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)
38. Lemaréchal, C.: In: Handbooks in Operations Research and Management Science, pp. 529–572. Elsevier (1989). https://doi.org/10.1016/s0927-0507(89)01008-x
39. Mäkelä, M.M., Karmitsa, N., Wilppu, O.: Multiobjective proximal bundle method for nonsmooth optimization. TUCS Technical Report No 1120, Turku Centre for Computer Science, Turku (2014)
40. Gebken, B., Peitz, S.: An efficient descent method for locally Lipschitz multiobjective optimization problems. J. Optim. Theory Appl. **80**, 3–29 (2021). https://doi.org/10.1007/s10957-020-01803-w
41. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002). https://doi.org/10.1109/4235.996017
42. Aliprantis, C., Border, K.: Infinite Dimensional Analysis: A Hitchhiker's Guide, 3rd edn. Springer-Verlag, Berlin (2006). https://doi.org/10.1007/3-540-29587-9
43. Brøndsted, A.: An Introduction to Convex Polytopes. Springer, New York (1983). https://doi.org/10.1007/978-1-4612-1148-8