Chapter 4
Mathematical Programming Problems with Vanishing Constraints

We study mathematical programming problems with vanishing constraints (MPVC) from the topological point of view. The critical point theory for MPVCs is presented. For that, we introduce the notion of a T-stationary point for the MPVC.

4.1 Applications and examples

We consider the mathematical programming problem with vanishing constraints (MPVC)

$$\text{MPVC:} \quad \min f(x) \text{ s.t. } x \in M[h,g,H,G], \tag{4.1}$$

with

$$M[h,g,H,G] := \{x \in \mathbb{R}^n \mid H_m(x) \geq 0, H_m(x)G_m(x) \leq 0, m = 1,\ldots,k, \\ h_i(x) = 0, i \in I, g_j(x) \geq 0, j \in J\},$$

where $h := (h_i, i \in I)^T \in C^2(\mathbb{R}^n, \mathbb{R}^{|I|})$, $g := (g_j, j \in J)^T \in C^2(\mathbb{R}^n, \mathbb{R}^{|J|})$, $H := (H_m, m = 1,\ldots,k)^T, G := (G_m, m = 1,\ldots,k)^T \in C^2(\mathbb{R}^n, \mathbb{R}^k)$, $f \in C^2(\mathbb{R}^n, \mathbb{R})$, $|I| \leq n$, $k \geq 0$, $|J| < \infty$. For simplicity, we write M for $M[h,g,H,G]$ if no confusion is possible.

The MPVC was introduced in [1] as a model for structural and topology optimization. It is motivated by the fact that the constraint G_m does not play any role whenever H_m is active. We refer the reader to [43, 44, 45, 46, 56, 55] for more details on optimality conditions, constraint qualifications, sensitivity, and numerical methods for the MPVC. Note that additional constraints $G_m(x) \geq 0$, $m = 1,\ldots,k$ would restrict the MPVC to a so-called mathematical program with complementarity constraints (MPCC). In addition to an MPCC feasible set, M is glued together from manifold pieces of **different dimensions** along their strata. Indeed, a typical MPVC feasible set

$$\mathbb{V} := \{(x,y) \,|\, x \geq 0, xy \geq 0\}$$

V. Shikhman, *Topological Aspects of Nonsmooth Optimization*, Springer Optimization and Its Applications 64, DOI 10.1007/978-1-4614-1897-9_4,

is depicted in Figure 21.

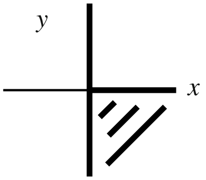

Figure 21 $\mathbb{V}$ **solution set of the basic vanishing constraint relation**

It represents the solution set of the basic vanishing constraint relations and exhibits one- and two-dimensional parts glued together at $(0,0)$.

Truss topology optimization

The following application of truss topology optimization is from [1]. The problem is to construct the optimal design of a truss structure. Let us consider a set of potential bars that are defined by the coordinates of their end nodes. For each potential bar, material parameters are given (Young's modulus E_i, relative moment of inertia s_i, and stress bounds $\sigma_i^t > 0$ and $\sigma_i^c < 0$ for tension and compression, respectively). These parameters are needed for the formulation of constraints to prevent structural failure when the potential bar is realized as a real bar. This the case if the calculated cross-sectional area a_i is positive. Finally, boundary conditions and external loads at some of the nodes are given. The problem now is to find cross-sectional areas a_i for each potential bar such that failure of the whole structure is prevented, the external load is carried by the structure, and a suitable objective function is minimal. The latter is usually the total weight of the structure or its deformation energy. In view of a practical realization of the calculated structure after optimization, one hopes that the optimal design will make use of only a few of the potential bars. Such behavior is typical in applied truss topology optimization problems. The main difficulty in formulating (and solving) the problem lies in the fact that constraints on structural failure can be formulated in a well-defined way only if there is some material giving mechanical response. However, most potential bars will possess a zero cross section at the optimizer. Hence, the truss topology optimization problem might be formulated as an MPVC:

$$\text{Truss-Top:} \quad \underset{(a,u)\in\mathbb{R}^M\times\mathbb{R}^d}{\text{minimize}} f(a,u) \text{ s.t.}$$

$$g(a,u) \leq 0, K(a)u = f^{\text{ext}},$$

$$a_i \geq 0, i = 1,\dots,M,$$

$$\sigma_i^c \leq \sigma(a,u) \leq \sigma_i^t \text{ if } a_i > 0, i = 1,\dots,M,$$

$$f_i^{\text{int}}(a,u) \geq f_i^{\text{buck}}(a,u) \text{ if } a_i > 0, i = 1,\dots,M.$$

Here, the vector $a \in \mathbb{R}^M$ contains the vector of cross-sectional areas of the potential bars and $u \in \mathbb{R}^d$ denotes the vector of nodal displacements of the structure under load, where d is the so-called degree of freedom of the structure, the number of free nodal displacement coordinates. The state variable u serves as an auxiliary variable. The objective function f expresses structural weight. The nonlinear system of equations $K(a)u = f^{\text{ext}}$ symbolizes force equilibrium of (given) external loads $f^{\text{ext}} \in \mathbb{R}^d$ and internal forces expressed via Hooke's law in terms of displacements and cross sections. The matrix $K(a) \in \mathbb{R}^{d\times d}$ is the global stiffness matrix corresponding to the structure a. The constraint $g(a,u) \leq 0$ is a resource constraint. If $a_i > 0$, then $\sigma_i(a,u) \in \mathbb{R}$ is the stress along the i-th bar. Similarly, if $a_i > 0$, $f_i^{\text{int}}(a,u) \in \mathbb{R}$ denotes the internal force along the i-th bar, and $f_i^{\text{buck}}(a)$ corresponds to the permitted Euler buckling force. Then the constraints on stresses and on local buckling make sense only if $a_i > 0$. Therefore, they must vanish from the problem if $a_i = 0$.

4.2 Critical point theory

Our goal is the investigation of the MPVC from a topological point of view. To this end, we introduce the new notion of a T-stationary point for the MPVC (see Definition 29). It turns out that the concept of T-stationarity is an adequate stationarity concept for topological considerations. In fact, we introduce the letter "T" for a stationarity concept that is **topologically** relevant rather than giving a tight first-order condition for local minimizers (see also the discussion below).

Furthermore, we study the behavior of the topological properties of lower-level sets

$$M^a := \{x \in M \,|\, f(x) \leq a\}$$

for the MPVC as the level $a \in \mathbb{R}$ varies. In particular, within this context, we present two basic theorems from Morse theory (see [63, 93] and Section A.1). First, we show that, for $a < b$, the set M^a is a strong deformation retract of M^b if the (compact) set

$$M_a^b := \{x \in M \,|\, a \leq f(x) \leq b\}$$

does not contain T-stationary points (see Theorem 40(a)). Second, if M_a^b contains exactly one (nondegenerate) T-stationary point, then M^b is shown to be homotopy-equivalent to M^a with a q-cell attached (see Theorem 40(b)). Here, the dimension q is the T-index (see Definitions 29 and 31). We refer the reader to [20] for details.

T-stationarity

Given $\bar{x} \in M$, we define the (active) index sets

$$J_0 = J_0(\bar{x}) := \{j \in J \,|\, g_j(\bar{x}) = 0\},$$

$$
\begin{aligned}
I_{0+} = I_{0+}(\bar{x}) &:= \{m \in \{1,\dots k\} \,|\, H_m(\bar{x}) = 0, G_m(\bar{x}) > 0\},\\
I_{0-} = I_{0-}(\bar{x}) &:= \{m \in \{1,\dots k\} \,|\, H_m(\bar{x}) = 0, G_m(\bar{x}) < 0\},\\
I_{+0} = I_{+0}(\bar{x}) &:= \{m \in \{1,\dots k\} \,|\, H_m(\bar{x}) > 0, G_m(\bar{x}) = 0\},\\
I_{00} = I_{00}(\bar{x}) &:= \{m \in \{1,\dots k\} \,|\, H_m(\bar{x}) = 0, G_m(\bar{x}) = 0\}.
\end{aligned}
$$

We call $J_0(\bar{x})$ the active inequality index set and $I_{00}(\bar{x})$ the biactive index set at $\bar{x}$. Note that, locally around $\bar{x}$, for $m \in I_{0+}$, the function H_m behaves like an ordinary equality constraint ($H_m(x) = 0$). For $m \in I_{0-}$ or $m \in I_{+0}$, the functions H_m and G_m behave locally like inequality constraints ($H_m(x) \geq 0$ or $G_m(x) \leq 0$, respectively).

Furthermore, we recall the well-known linear independence constraint qualification (LICQ) for the MPVC (e.g., [1]), which is said to hold at $\bar{x} \in M$ if the vectors

$$
\begin{aligned}
&D^T h_i(\bar{x}),\ i \in I, D^T H_m(\bar{x}),\ m \in I_{0+},\\
&D^T g_j(\bar{x}),\ j \in J_0,\ D^T H_m(\bar{x}),\ m \in I_{0-},\ D^T G_m(\bar{x}),\ m \in I_{+0},\\
&D^T H_m(\bar{x}),\ D^T G_m(\bar{x}),\ m \in I_{00}
\end{aligned}
$$

are linearly independent.

We introduce the notion of a T-stationary point, which is crucial for the following.

Definition 29 (T-stationary point). A point $\bar{x} \in M$ is called T-stationary for the MPVC if there exist real numbers $\bar{\lambda}_i$, $i \in I$, $\bar{\alpha}_m$, $m \in I_{0+}$, $\bar{\mu}_j$, $j \in J_0$, $\bar{\beta}_m$, $m \in I_{0-}$, $\bar{\gamma}_m$, $m \in I_{+0}$, $\bar{\delta}^H_m$, $\bar{\delta}^G_m$, $m \in I_{00}$ (Lagrange multipliers) such that

$$
\begin{aligned}
Df(\bar{x}) = &\sum_{i \in I} \bar{\lambda}_i Dh_i(\bar{x}) + \sum_{m \in I_{0+}} \bar{\alpha}_m DH_m(\bar{x})\\
&+ \sum_{j \in J_0} \bar{\mu}_j Dg_j(\bar{x}) + \sum_{m \in I_{0-}} \bar{\beta}_m DH_m(\bar{x}) + \sum_{m \in I_{+0}} \bar{\gamma}_m DG_m(\bar{x})\\
&+ \sum_{m \in I_{00}} \left(\bar{\delta}^H_m DH_m(\bar{x}) + \bar{\delta}^G_m DG_m(\bar{x}) \right),
\end{aligned} \tag{4.2}
$$

$$\bar{\mu}_j \geq 0 \text{ for all } j \in J_0, \tag{4.3}$$

$$\bar{\beta}_m \geq 0 \text{ for all } m \in I_{0-}, \tag{4.4}$$

$$\bar{\gamma}_m \leq 0 \text{ for all } m \in I_{+0}, \tag{4.5}$$

$$\bar{\delta}^G_m \leq 0 \text{ and } \bar{\delta}^H_m \cdot \bar{\delta}^G_m \geq 0 \text{ for all } m \in I_{00}. \tag{4.6}$$

In the case where the LICQ holds at $\bar{x} \in M$, the Lagrange multipliers in (4.2) are uniquely determined.

Given a T-stationary point $\bar{x} \in M$ for the MPVC, we set

$$
\begin{aligned}
M(\bar{x}) := \{x \in \mathbb{R}^n \mid\ &h_i(x) = 0,\ i \in I, H_m(x) = 0,\ m \in I_{0+},\ g_j(x) = 0,\ j \in J_0,\\
&H_m(x) = 0,\ m \in I_{0-},\ G_m(x) = 0,\ m \in I_{+0},\\
&H_m(x) = 0,\ G_m(x) = 0,\ m \in I_{00}\}.
\end{aligned}
$$

Obviously, $M(\bar{x}) \subset M$ and, in the case where the LICQ holds at $\bar{x}$, $M(\bar{x})$ is locally at $\bar{x}$ a C^2-manifold.

Definition 30 (Nondegenerate T-stationary point). A T-stationary point $\bar{x} \in M$ with Lagrange multipliers as in Definition 29 is called nondegenerate if the following conditions are satisfied:

ND1: LICQ holds at $\bar{x}$.
ND2: $\bar{\mu}_j > 0$ for all $j \in J_0$, $\bar{\beta}_m > 0$ for all $m \in I_{0-}$, $\bar{\gamma}_m < 0$ for all $m \in I_{+0}$.
ND3: $D^2L(\bar{x}) \mid_{T_{\bar{x}}M(\bar{x})}$ is nonsingular.
ND4: $\bar{\delta}_m^H < 0$ and $\bar{\delta}_m^G < 0$ for all $m \in I_{00}$.

Here, the matrix D^2L stands for the Hessian of the Lagrange function L,

$$\begin{aligned} L(x) := f(x) - \sum_{i\in I}\bar{\lambda}_i h_i(x) - \sum_{m\in I_{0+}} \bar{\alpha}_m H_m(x) \\ - \sum_{j\in J_0} \bar{\mu}_j g_j(x) - \sum_{m\in I_{0-}} \bar{\beta}_m H_m(x) - \sum_{m\in I_{+0}} \bar{\gamma}_m G_m(x) \\ - \sum_{m\in I_{00}} \left(\bar{\delta}_m^H H_m(x) - \bar{\delta}_m^G G_m(x)\right), \end{aligned} \tag{4.7}$$

and $T_{\bar{x}}M(\bar{x})$ denotes the tangent space of $M(\bar{x})$ at $\bar{x}$,

$$\begin{aligned} T_{\bar{x}}M(\bar{x}) := \{\xi \in \mathbb{R}^n \mid\ & Dh_i(\bar{x})\xi = 0, i \in I, \\ & DH_m(\bar{x})\xi = 0, m \in I_{0+}, \\ & Dg_j(\bar{x})\xi = 0, j \in J_0, \\ & DH_m(\bar{x})\xi = 0, m \in I_{0-}, \\ & DG_m(\bar{x})\xi = 0, m \in I_{+0}, \\ & DH_m(\bar{x})\xi = 0, DG_m(\bar{x})\xi = 0, m \in I_{00}\}. \end{aligned}$$

Condition ND3 means that the matrix $V^T D^2L(\bar{x})V$ is nonsingular, where V is some matrix whose columns form a basis for the tangent space $T_{\bar{x}}M(\bar{x})$.

Definition 31 (T-index). Let $\bar{x} \in M$ be a nondegenerate T-stationary point with Lagrange multipliers as in Definition 30. The number of negative eigenvalues of $D^2L(\bar{x}) \mid_{T_{\bar{x}}M(\bar{x})}$ in ND3 is called the quadratic index (QI) of $\bar{x}$. The number of negative pairs $(\bar{\delta}_m^H, \bar{\delta}_m^G)$, $m \in I_{00}$ in ND4 equals $|I_{00}|$ and is called the biactive index (BI) of $\bar{x}$. The number $(QI + BI)$ is called the T-index of $\bar{x}$.

Note that in the absence of biactive vanishing constraints, the T-index has only the QI part and coincides with the well-known quadratic index of a nondegenerate Karush-Kuhn-Tucker point in nonlinear programming or, equivalently, with the Morse index (see [63, 83, 93] and Section 1.4). Also note that the biactive index BI is completely determined by the cardinality of I_{00}, in contrast to, for example, the biactive index for MPCCs as defined in Section 2.3 (see also [69]).

The following proposition uses the T-index for the characterization of a local minimizer.

Proposition 8. *(i) Assume that $\bar{x}$ is a local minimizer for the MPVC and that the LICQ holds at $\bar{x}$. Then, $\bar{x}$ is a T-stationary point for the MPVC.*

(ii) Let $\bar{x}$ be a nondegenerate T-stationary point for the MPVC. Then, $\bar{x}$ is a local minimizer for the MPVC if and only if its T-index is equal to zero.

Proof. (i) From [1] it is known that under the LICQ a local minimizer $\bar{x}$ for the MPVC is a strongly stationary point,meaning (4.2)–(4.5) hold and

$$\bar{\delta}_m^G = 0 \text{ and } \bar{\delta}_m^H \geq 0 \text{ for all } m \in I_{00}. \tag{4.8}$$

Clearly, a strongly stationary point is T-stationary as well.

(ii) Let $\bar{x}$ be a nondegenerate T-stationary local minimizer for the MPVC. As in (i), we claim that $\bar{x}$ is also strongly stationary. Comparing ND4 and (4.8), we see that $BI = |I_{00}| = 0$. Then, locally around $\bar{x}$, the MPVC behave like an ordinary nonlinear program, and using standard results on the quadratic index, we obtain that $QI = 0$. The other direction is trivial. □

The next genericity and stability results justify the LICQ assumption as well as the introduction of nondegeneracy for T-stationary points in the MPVC.

Theorem 38 (Genericity and Stability).

(i) Let $\mathscr{F}$ denote the subset of

$$C^2(\mathbb{R}^n, \mathbb{R}^{|I|}) \times C^2(\mathbb{R}^n, \mathbb{R}^{|J|}) \times C^2(\mathbb{R}^n, \mathbb{R}^k) \times C^2(\mathbb{R}^n, \mathbb{R}^k)$$

consisting of those (h, g, H, G) for which the LICQ holds at all points $x \in M[h, g, H, G]$. Then, $\mathscr{F}$ is C_s^2-open and -dense.

(ii) Let $\mathscr{D}$ denote the subset of

$$C^2(\mathbb{R}^n, \mathbb{R}) \times C^2(\mathbb{R}^n, \mathbb{R}^{|I|}) \times C^2(\mathbb{R}^n, \mathbb{R}^{|J|}) \times C^2(\mathbb{R}^n, \mathbb{R}^k) \times C^2(\mathbb{R}^n, \mathbb{R}^k)$$

consisting of those problem data (f, h, g, H, G) for which each T-stationary point is nondegenerate. Then, $\mathscr{D}$ is C_s^2-open and -dense.

Proof. (i) We define the set

$$M_{\text{DISJ}} := \{x \in \mathbb{R}^n \mid \max\{H_m(x), G_m(x)\} \geq 0, m = 1, \ldots, k, \\ h_j(x) = 0, i \in I, g_j(x) \geq 0, j \in J\}.$$

M_{DISJ} is the feasible set of a disjunctive optimization problem (see [71]). We obtain from the corresponding results on disjunctive optimization that the subset of problem data for which the LICQ holds for all $x \in M_{\text{DISJ}}$ is C_s^2-dense and C_s^2-open (see [71], Lemmas 2.4 and 2.5). Recalling that the notions of the LICQ for disjunctive optimization problems and MPVCs are the same, and that M is a subset of M_{DISJ}, the desired result follows immediately.

(ii) The proof is based on the application of the jet transversality theorem, for details, see, for example, [63] and Section B.2. For subsets $\tilde{J} \subseteq J$ and $\tilde{H}, \tilde{G} \subseteq \{1, \ldots, k\}$, and sets $D_{\tilde{J}} \subseteq \tilde{J}$, $D_{\tilde{H}} \subseteq \tilde{H}$, and $D_{\tilde{G}} \subseteq \tilde{G}$ and $r \in \{0, \ldots, \dim(T_{\bar{x}}M(\bar{x}))\}$, we consider the set Γ of x such that the following conditions are satisfied:

(m1) $g_j(x) = H_i(x) = G_l(x) = 0$ for all $j \in \tilde{J}$, $i \in \tilde{H}$, $l \in \tilde{G}$.

(m2) $$Df(x) \in \text{span} \left\{ \begin{array}{l} Dg_j(x),\, j \in \tilde{J} \setminus D_{\tilde{J}}, \\ DH_i(x),\, i \in \tilde{H} \setminus D_{\tilde{H}}, \\ DG_l(x),\, l \in \tilde{G} \setminus D_{\tilde{G}} \end{array} \right\}.$$

(m3) The matrix $D^2L(x)|_{T_{\bar{x}}M(\bar{x})}$ has rank r.

Now it suffices to show that Γ is generically empty whenever one of the sets $D_{\tilde{J}}$, $D_{\tilde{H}}$, or $D_{\tilde{G}}$ is nonempty or the rank r of the matrix in (m3) is not full. This would mean, respectively, that a Lagrange multiplier in the equality (4.2) vanishes (see ND2, ND4) or the rank condition ND3 fails to hold.

In fact, the available degrees of freedom of the variables involved in Γ are n. The loss of freedom caused by (m1) is at least $d := |\tilde{J}| + |\tilde{H}| + |\tilde{G}|$, and the loss of freedom caused by (m2) is at least (supposing that the gradients on the right-hand side are linearly independent (ND1) and the sets $D_{\tilde{J}}, D_{\tilde{H}}, D_{\tilde{G}}$ are empty) $n-d$. Hence, the total loss of freedom is n. We conclude that a further nondegeneracy would exceed the total available degrees of freedom n. By virtue of the jet transversality theorem, generically the set Γ must be empty.

For the openness result, we can argue in a standard way (see, for example, [63]). Locally, T-stationarity can be rewritten via stable equations. Then, the implicit function theorem for Banach spaces can be applied to follow nondegenerate T-stationary points w.r.t. (local) C^2-perturbations of defining functions. Then a standard globalization procedure exploiting the specific properties of the strong C^2-topology can be used to construct a (global) C^2_s-neighborhood of problem data for which the nondegeneracy property is stable.□

Morse lemma for the MPVC

For the proof of the results mentioned above we locally describe the MPVC feasible set under the LICQ (see Lemma 24). Moreover, an equivariant Morse lemma for the MPVC is derived in order to obtain suitable normal forms for the objective function at nondegenerate T-stationary points (see Theorem 39).

Without loss of generality, we assume that at the particular point of interest $\bar{x} \in M$ it holds that

$$J_0 = \{1, \dots, |J_0|\},$$

$$I_{0+} = \{1, \dots, |I_{0+}|\},$$

$$I_{0-} = \{|I_{0+}|+1, \dots, |I_{0+}|+|I_{0-}|\},$$

$$I_{+0} = \{|I_{0+}|+|I_{0-}|+1, \dots, |I_{0+}|+|I_{0-}|+|I_{+0}|\},$$

$$I_{00} = \{|I_{0+}|+|I_{0-}|+|I_{+0}|+1, \dots, |I_{0+}|+|I_{0-}|+|I_{+0}|+|I_{00}|\}.$$

We put $s := |I| + |I_{0+}|$, $r := s + |J_0| + |I_{0-}|$, $q := r + |I_{+0}|$, $p := n - q - 2\,|I_{00}|$.

For the proof of Theorem 40, we need a local description of the MPVC feasible set under the LICQ.

Definition 32. The feasible set M admits a local C^r-coordinate system of $\mathbb{R}^n$ ($r \geq 1$) at $\bar{x}$ by means of a C^r-diffeomorphism $\Phi : U \longrightarrow V$ with open $\mathbb{R}^n$-neighborhoods U and V of $\bar{x}$ and 0, respectively, if it holds that

(i) $\Phi(\bar{x}) = 0$,

(ii) $\Phi(M \cap U) = \left(\{0_s\} \times \mathbb{H}^{|J_0|+|I_{0-}|} \times (-\mathbb{H})^{|I_{+0}|} \times \mathbb{V}^{|I_{00}|} \times \mathbb{R}^p\right) \cap V.$

Lemma 24. *Suppose that the LICQ holds at $\bar{x} \in M$. Then M admits a local C^2-coordinate system of $\mathbb{R}^n$ at $\bar{x}$.*

Proof. Choose vectors $\xi_l \in \mathbb{R}^n$, $l = 1, \ldots, p$, which form, together with the vectors

$$\begin{aligned}
&D^T h_i(\bar{x}),\ i \in I, D^T H_m(\bar{x}),\ m \in I_{0+},\\
&D^T g_j(\bar{x}),\ j \in J_0,\ D^T H_m(\bar{x}),\ m \in I_{0-},\ D^T G_m(\bar{x}),\ m \in I_{+0},\\
&D^T H_m(\bar{x}), D^T G_m(\bar{x}),\ m \in I_{00},
\end{aligned}$$

a basis for $\mathbb{R}^n$. Next we put

$$\left.\begin{array}{ll}
y_i & := h_i(x),\ i \in I,\\
y_{|I|+m} & := H_m(x),\ m \in I_{0+},\\
y_{|I|+|I_{0+}|+j} & := g_j(x),\ j \in J_0,\\
y_{|I|+|J_0|+m} & := H_m(x),\ m \in I_{0-},\\
y_{|I|+|J_0|+m} & := G_m(x),\ m \in I_{+0},\\
y_{|I|+|J_0|+2m-1} & := H_m(x),\ m \in I_{00},\\
y_{|I|+|J_0|+2m} & := G_m(x),\ m \in I_{00},\\
y_{n-p+l} & := \xi_l^T(x - \bar{x}),\ l = 1, \ldots, p
\end{array}\right\} \tag{4.9}$$

or, for short,

$$y = \Phi(x). \tag{4.10}$$

Note that $\Phi \in C^2(\mathbb{R}^n, \mathbb{R}^n)$, $\Phi(\bar{x}) = 0$, and the Jacobian matrix $D\Phi(\bar{x})$ is nonsingular (by virtue of the LICQ and the choice of ξ_l, $l = 1, \ldots, p$). By means of the implicit function theorem, there exist open neighborhoods U of $\bar{x}$ and V of 0 such that $\Phi : U \longrightarrow V$ is a C^2-diffeomorphism. By shrinking U if necessary, we can guarantee that $J_0(x) \subset J_0$, $I_{0-}(x) \subset I_{0-}$, $I_{+0}(x) \subset I_{+0}$ and $I_{00}(x) \subset I_{00}$ for all $x \in M \cap U$. Thus, property (ii) in Definition 32 follows directly from the definition of Φ. □

Definition 33. We will refer to the C^2-diffeomorphism Φ defined by (4.9) and (4.10) as the standard diffeomorphism.

Remark 29. It follows from the proof of Lemma 24 that the Lagrange multipliers at a nondegenerate T-stationary point are the corresponding partial derivatives of the objective function in new coordinates given by the standard diffeomorphism (see [65], Lemma 2.2.1). Moreover, the Hessian with respect to the last p coordinates corresponds to the restriction of the Lagrange function's Hessian on the respective tangent space (cf. [65], Lemma 2.2.10).

We derive an equivariant Morse lemma for the MPVC in order to obtain suitable normal forms for the objective function at nondegenerate T-stationary points.

Theorem 39 (Morse lemma for MPVC). *Suppose that $\bar{x}$ is a nondegenerate T-stationary point for the MPVC with quadratic index QI, biactive index BI, and T-index = QI + BI. Then, there exists a local C^1-coordinate system $\Psi : U \longrightarrow V$ of $\mathbb{R}^n$ around $\bar{x}$ (according to Definition 32) such that*

$$f \circ \Psi^{-1}(0_s, y_{s+1}, \dots, y_n) =$$

$$f(\bar{x}) + \sum_{i=1}^{|J_0|+|I_{0-}|} y_{i+s} - \sum_{j=1}^{|I_{+0}|} y_{j+r} - \sum_{m=1}^{|I_{00}|} \left(y_{2j-1+q} + y_{2j+q}\right) + \sum_{k=1}^{p} \pm y^2_{k+n-p}, \qquad (4.11)$$

where $y \in \{0_s\} \times \mathbb{H}^{|J_0|+|I_{0-}|} \times (-\mathbb{H})^{|I_{+0}|} \times \mathbb{V}^{|I_{00}|} \times \mathbb{R}^p$. Moreover, in (4.11) there are exactly $BI = |I_{00}|$ negative linear pairs and QI negative squares.

Proof. Without loss of generality, we may assume $f(\bar{x}) = 0$. Let $\Phi : U \longrightarrow V$ be a standard diffeomorphism according to Definition 33. We put $\bar{f} := f \circ \Phi^{-1}$ on the set $\left(\{0_s\} \times \mathbb{H}^{|J_0|+|I_{0-}|} \times (-\mathbb{H})^{|I_{+0}|} \times \mathbb{V}^{|I_{00}|} \times \mathbb{R}^p\right) \cap V$. We may assume $s = 0$ from now on. In view of Remark 29, we have at the origin

(i) $\dfrac{\partial \bar{f}}{\partial y_i} > 0,\ i = 1, \dots, |J_0| + |I_{0-}|,$

(ii) $\dfrac{\partial \bar{f}}{\partial y_{j+r}} < 0,\ j = 1, \dots, |I_{+0}|,$

(iii) $\dfrac{\partial \bar{f}}{\partial y_{2m-1+q}} < 0$ and $\dfrac{\partial \bar{f}}{\partial y_{2m+q}} < 0$ for exactly BI indices $m = 1, \dots, |I_{00}|,$

(iv) $\dfrac{\partial \bar{f}}{\partial y_{k+n-p}} = 0,\ k = 1, \dots, p$ and $\left(\dfrac{\partial^2 \bar{f}}{\partial y_{k_1+n-p} \partial y_{k_2+n-p}}\right)_{1 \le k_1, k_2 \le p}$ is a nonsingular matrix with QI negative eigenvalues.

We denote $\bar{f}$ by f. Under the following coordinate transformations the set $\mathbb{H}^{|J_0|+|I_{0-}|} \times (-\mathbb{H})^{|I_{+0}|} \times \mathbb{V}^{|I_{00}|} \times \mathbb{R}^p$ will be transformed in itself (equivariance). As an abbreviation, we put $y = (Y_{n-p}, Y^p)$, where $Y_{n-p} = (y_1, \dots, y_{n-p})$ and $Y^p = (y_{n-p+1}, \dots, y_n)$. We write

$$f(Y_{n-p}, Y^p) = f(0, Y^p) + \int_0^1 \frac{d}{dt} f(tY_{n-p}, Y^p)dt = f(0, Y^p) + \sum_{i=1}^{n-p} y_i d_i(y),$$

where $d_i \in C^1$, $i = 1, \dots, n-p$.

In view of (iv), we may apply the Morse lemma on the C^2-function $f(0, Y^p)$ (see Theorem 2.8.2 of [63]) without affecting the coordinates Y_{n-p}. The corresponding coordinate transformation is of class C^1. Denoting the transformed functions f, d_j again by f, d_j, we obtain

$$f(y) = \sum_{i=1}^{n-p} y_i d_i(y) + \sum_{k=1}^{p} \pm y^2_{k+n-p}.$$

Note that $d_i(0) = \dfrac{\partial f}{\partial y_i}(0)$, $i = 1,\dots,n-p$. Recalling (i)–(iii), we have

$$y_i|d_i(y)|,\ i = 1,\dots,n-p, \quad y_j,\ j = n-p+1,\dots,n, \tag{4.12}$$

as new local C^1-coordinates. Denoting the transformed function f again by f and recalling the signs in (i)–(iii), we obtain (4.11). Here, the coordinate transformation Ψ is understood as the composite of all previous ones.□

Deformation and Cell-Attachment

We state and prove the main deformation and cell-attachment theorems for the MPVC. Recall that for $a, b \in \mathbb{R}$, $a < b$, the sets M^a and M_a^b are defined as

$$M^a := \{x \in M \,|\, f(x) \leq a\}$$

and

$$M_a^b := \{x \in M \,|\, a \leq f(x) \leq b\}.$$

Theorem 40. *Let M_a^b be compact, and suppose that the LICQ is satisfied at all points $x \in M_a^b$.*

(a) **(Deformation theorem)** *If M_a^b does not contain any T-stationary point for the MPVC, then M^a is a strong deformation retract of M^b.*

(b) **(Cell-attachment theorem)** *If M_a^b contains exactly one (nondegenerate) T-stationary point for the MPVC, say $\bar{x}$, and if $a < f(\bar{x}) < b$ and the T-index of $\bar{x}$ is equal to q, then M^b is homotopy-equivalent to M^a with a q-cell attached.*

Proof. (a) Let $\bar{x} \in M_a^b$. After a coordinate transformation with the standard diffeomorphism from Definition 32 and Remark 29, we may assume that $\bar{x} = 0$ and locally $M = \{0_s\} \times \mathbb{H}^{|J_0|+|I_{0-}|} \times (-\mathbb{H})^{|I_{+0}|} \times \mathbb{V}^{|I_{00}|} \times \mathbb{R}^p$. From Remark 29 and the fact that $\bar{x}$ is not a T-stationary point (see Definition 29), one of the following cases holds:

(a) There exists $j \in \{1,\dots,p\}$ with $\dfrac{\partial f}{\partial y_{n-p+j}}(0) \neq 0$.

(b) There exists $j \in \{1,\dots,|J_0| + |I_{0-}|\}$ with $\dfrac{\partial f}{\partial y_{s+j}}(0) < 0$.

(c) There exists $j \in \{1,\dots,|I_{+0}|\}$ with $\dfrac{\partial f}{\partial y_{r+j}}(0) > 0$.

(d) There exists $m \in I_{00}$ with $\dfrac{\partial f}{\partial y_{q+2m}}(0) > 0$.

(e) There exists $m \in I_{00}$ with $\dfrac{\partial f}{\partial y_{q+2m-1}}(0) > 0$ and $\dfrac{\partial f}{\partial y_{q+2m}}(0) < 0$.

We set

$$D := \{x \in M_a^b \mid \text{one of cases a)–d) holds}\} \quad \text{and} \quad L := M_a^b \setminus D.$$

The proof consists of the local argument and its globalization.

Local argument. We prove that for each $\bar{x} \in M_a^b$ there exists an $\mathbb{R}^n$-neighborhood $\mathscr{U}_{\bar{x}}$ of $\bar{x}$, a $t_{\bar{x}} > 0$, and a flow

$$\Psi^{\bar{x}} : [0,t_{\bar{x}}) \times M^b \cap \mathscr{U}_{\bar{x}} \to M, \ (t,x) \mapsto \Psi^{\bar{x}}(t,x),$$

with:

1. $\Psi^{\bar{x}}(0,x) = x$ for all $x \in M^b \cap \mathscr{U}_{\bar{x}}$.
2. $\Psi^{\bar{x}}(t_2, \Psi^{\bar{x}}(t_1,x)) = \Psi^{\bar{x}}(t_1+t_2,x)$ for all $x \in M^b \cap \mathscr{U}_{\bar{x}}$ and $t_1,t_2 \geq 0$ with $t_1+t_2 \in [0,t_{\bar{x}})$.
3. $f(\Psi^{\bar{x}}(t,x)) \leq f(x) - t$ for all $x \in M^b \cap \mathscr{U}_{\bar{x}}$ and $t \in [0,t_{\bar{x}})$.
4. If $\bar{x} \in D$, then $\Psi^{\bar{x}}$ is a C^2-flow corresponding to a C^1-vector field. If $\bar{x} \in L$, then $\Psi^{\bar{x}}$ is a Lipschitz flow.

We consider the constructions of the local flows in Cases a)–e).

Cases (a)–(c). We can use standard methods to construct a local flow induced by a C^1-vector field. To see this, note that the behavior of partial derivatives in Cases (a)–(c) give us a descent direction that—due to the structure of M in local coordinates—is feasible for $t_{\bar{x}} > 0$. (This is a standard construction for generalized manifolds with boundary; see Theorems 2.7.6 and 3.2.26 of [63] for details and also the proof of Theorem 20).

If the violation of T-stationarity is exclusively due to the coordinates belonging to the set $\mathbb{V}^{|I_{00}|}$ (i.e. one of the cases (d) and (e) holds), we have to construct a new flow.

Case (d). Using an (additional) local coordinate transformation leaving M invariant, analogous to the proof of Theorem 39, we obtain

$$f(y) = y_{q+2m} + f(y_1,\ldots,y_{q+2m-1},0,y_{q+2m+1},\ldots,y_n).$$

We define a local vector field as $\tilde{F}^{\bar{x}}(y) := (0,\ldots,0,-1,0,\ldots,0)^T$. After the inverse change of local coordinates, $\tilde{F}^{\bar{x}}$ induces the flow, which fits the local argument.

Case e). Again, as in the proof of Theorem 39, we may assume that

$$f(y) = y_{q+2m-1} - y_{q+2m} + f(y_1,\ldots,y_{q+2m-2},0,0,y_{q+2m+1},\ldots,y_n).$$

We define a two-dimensional flow $\Phi(t,z)$ for $z = (z_1,z_2) \in \mathbb{V}$ as

$$\Phi(t,z_1,z_2) := \begin{cases} \begin{pmatrix} \max\left\{0, \left(1 - \frac{t}{z_1-z_2}\right) \cdot z_1\right\} \\ \left[\left(1 - \frac{t}{z_1-z_2}\right) \cdot z_2\right]^- + [t-(z_1-z_2)]^+ \end{pmatrix} & \text{for } z_2 < 0, \\ \\ \begin{pmatrix} 0 \\ t-(z_1-z_2) \end{pmatrix} & \text{for } z_2 \geq 0. \end{cases}$$

Here, $[\cdot]^-$ is the negative and $[\cdot]^+$ the positive part of a real number.

Note that the flow Φ is Lipschitz on $\mathbb{R} \times \mathbb{V}$. Moreover, due to the definition of Φ, we get that the flow $\Psi^{\bar{x}}$ defined (again in new coordinates) by

$$\Psi_i(y) := \begin{cases} y_i & \text{for } i \in \{1,\dots,n\} \setminus \{q+2m-1, q+2m\}, \\ \Phi_1(y_{q+2m-1}) & \text{for } i = q+2m-1, \\ \Phi_2(y_{q+2m}) & \text{for } i = q+2m. \end{cases}$$

fits the local argument. Here, Ψ_i and Φ_i stands for the i-th components of Ψ and Φ, respectively.

Globalization. Now we construct a global flow Ψ on M_a^b. Suppose for a moment that there exists a flow Ψ_L on a neighborhood $\mathscr{U}_L$ of L with the properties (i) to (iv). We choose a smaller neighborhood $\mathscr{W}_L$ of L such that the closure $\overline{\mathscr{W}_L}$ of $\mathscr{W}_L$ is contained in $\mathscr{U}_L$. Furthermore, we choose an arbitrary open covering $\{\mathscr{U}_x \mid x \in M_a^b \setminus \mathscr{U}_L\}$ of $M_a^b \setminus \mathscr{U}_L$ induced by the domains of the C^2-flows corresponding to cases (a)–(d). Since $M_a^b \setminus \mathscr{U}_L$ is compact we find a finite subcovering $\{\mathscr{U}_x \mid x \in \bar{D}\}$. Here $\bar{D}$ is a finite subset of D. Without loss of generality, we may assume that for all $x \in \bar{D}$ the closure $\overline{\mathscr{U}_x}$ of $\mathscr{U}_x$ is disjoint with $\overline{\mathscr{W}_L}$. By construction, it holds that $\{\mathscr{U}_x \mid x \in \bar{D}\} \cup (\mathscr{U}_L \setminus \overline{\mathscr{W}_L})$ is a finite open covering of $M_a^b \setminus \mathscr{W}_L$. The crucial argument is now that outside the set L the flow Ψ_L is induced by a C^1-vector field. (Note that Φ only has a singularity for $t = z_1 - z_2$.) Therefore, we can construct a flow on $M_a^b \setminus \overline{\mathscr{W}_L}$ by using a C^∞-partition of unity subordinate to the open covering $\{\mathscr{U}_x \mid x \in \bar{D}\} \cup (\mathscr{U}_L \setminus \overline{\mathscr{W}_L})$. This enables us to construct a global C^1-vector field. The flow Ψ_D obtained by integration fulfills the desired properties. (See Theorem 3.3.14 of [63] for details on this procedure.) By construction, Ψ_L and Ψ_D can be glued together into one flow Ψ on M_a^b.

We obtain for $x \in M_a^b$ a unique $t_a(x) > 0$ with $\Psi(t_a(x), x) \in M^a$ from the properties of Ψ (which are induced by local properties of the flows Ψ^x). It is not hard (but technical) to realize that $t_a : x \mapsto t_a(x)$ is Lipschitz. Finally, we define $r : [0,1] \times M^b \to M^b$ as

$$r(\tau, x) := \begin{cases} x & \text{for } x \in M^a,\ \tau \in [0,1], \\ \Psi(\tau \cdot t_a(x), x) & \text{for } x \in M_a^b,\ \tau \in [0,1]. \end{cases}$$

The mapping r provides that M^a is a strong deformation retract of M^b.

It remains to construct the flow Ψ_L. Since this construction is highly technical, we only present a short outline. The main idea is to construct the flow along strata inside L; here the strata are induced by all possible subsets of active constraints $H_1, G_1, \dots, H_m, G_m$. Along a given stratum, we find a differentiable family of standard coordinate systems (see Lemma 24). This enables us to define a flow along this stratum just by applying flows like Φ on fixed components that depend on the coordinate system. By introducing notions of a distance from a point in the embedding space to the strata, we can construct homotopies (via Lipschitz continuous time scaling) between the different branches of the stratification and the corresponding flows. (For details on such constructions with the aid of tube systems, we refer to [27].)

(b) From the deformation theorem (Theorem 40(a)), we may assume that, w.l.o.g., a and b are small enough that we can work in local coordinates. Therefore, we consider the normal form (2.19) from Theorem 39,

$$f(y) = \sum_{i=1}^{|J_0|+|I_{0-}|} y_{s+i} - \sum_{j=1}^{|I_{+0}|} y_{r+j} - \sum_{m=1}^{|I_{00}|} (y_{q+2m-1} + y_{q+2m}) + \sum_{l=1}^{p} \pm y_{n-p+l}^2,$$

with $y \in M := \{0_s\} \times \mathbb{H}^{|J_0|+|I_{0-}|} \times (-\mathbb{H})^{|I_{+0}|} \times \mathbb{V}^{|I_{00}|} \times \mathbb{R}^p$.

We set

$$M_{\text{MPCC}} := \{0_s\} \times \mathbb{H}^{|J_0|+|I_{0-}|} \times (-\mathbb{H})^{|I_{+0}|} \times (\partial \mathbb{H}^2)^{|I_{00}|} \times \mathbb{R}^p.$$

Note that M_{MPCC} differs from M by the appearance of $(\partial \mathbb{H}^2)^{|I_{00}|}$ instead of $\mathbb{V}^{|I_{00}|}$.

For $c \in \mathbb{R}$, it holds that $M^c_{\text{MPCC}} := \{y \in M_{\text{MPCC}} \mid f(y) \le c\}$ is a strong deformation retract of $M^c := \{y \in M \mid f(y) \le c\}$. In fact, we define a mapping $g : M^c \to M^c_{\text{MPCC}}$ with

$$y_i \mapsto \begin{cases} 0 & i \in \{q+2m \mid m = 1, \dots, |I_{00}|\} \text{ and } y_i < 0, \\ y_i & \text{else.} \end{cases}$$

We see that there is a (convex combination) homotopy between g and the identity on M^c. If $(y_{q+2m-1}, y_{q+2m}) \in \mathbb{V}$, then $(y_{q+2m-1}, 0) \in \partial \mathbb{H}^2$ and, moreover, $f(g(y)) \le f(y)$ for all $y \in M^c$ (i.e., g in fact maps to M^c_{MPCC}). Hence, M^c_{MPCC} is a strong deformation retract of M^c.

According to Definition 11, it holds that $\bar{y} = 0$ is a nondegenerate C-stationary point of the MPCC defined by f and the set M_{MPCC}. Since $\bar{y} = 0$ is the only C-stationary point, Theorem 20(b) implies that M^b_{MPCC} is homotopy-equivalent to M^a_{MPCC} with a $\tilde{q}$-cell attached. Note that $\tilde{q}$ is the so-called C-index for the corresponding MPCC. Here, we have that the C-index $\tilde{q}$ w.r.t. the MPCC coincides with the T-index q w.r.t. the MPVC. Hence

$$M^b_{\text{MPCC}} \simeq (M^a_{\text{MPCC}} \text{ with a } q\text{-cell attached}).$$

We know from the considerations above that M^c is homotopy-equivalent to M^c_{MPCC} for $c = a, b$. Furthermore, we note that the cell attachment on a homotopy-equivalent space is induced via the corresponding homotopy mapping. Finally, using the fact that homotopy equivalence is an equivalence relation, we obtain that M^b is homotopy-equivalent to M^a with a q-cell attached.□

Different stationarity concepts

We briefly review well-known definitions of various stationarity concepts and connections between them (see [1, 43, 44, 45, 46, 56]).

Definition 34. Let $\bar{x} \in M$.

(i) $\bar{x}$ is called weakly stationary if (4.2)–(4.5) hold and

$$\bar{\delta}^G_m \leq 0 \text{ for all } m \in I_{00}.$$

(ii) $\bar{x}$ is called M-stationary if (4.2)–(4.5) hold and

$$\bar{\delta}^G_m \leq 0 \text{ and } \bar{\delta}^G_m \cdot \bar{\delta}^H_m = 0 \text{ for all } m \in I_{00}.$$

(iii) $\bar{x}$ is called strongly stationary if (4.2)–(4.5) hold and

$$\bar{\delta}^G_m = 0 \text{ and } \bar{\delta}^H_m \geq 0 \text{ for all } m \in I_{00}.$$

Note that a strongly stationary point is M-stationary and the latter is T-stationary. We see that M- and strongly stationary points describe local minima tighter than T-stationary points. Moreover, strong stationarity is the tightest condition for a local minimizer under the LICQ. It is worth mentioning that M-stationarity exhibits a full calculus in the sense of Mordukhovich (see [94]). The scheme in Figure 22 illustrates the stationarity concepts above.

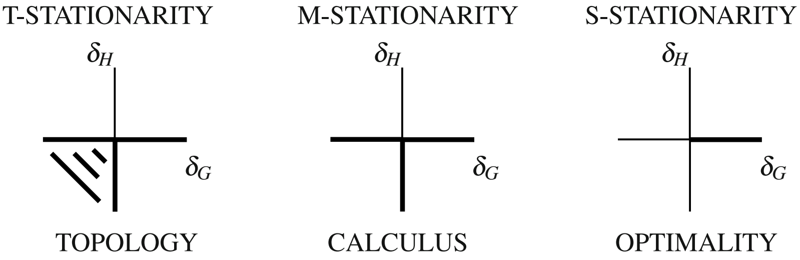

Figure 22 Stationarity concepts in MPVC

However, M- and strong stationarity exclude T-stationary points with $BI > 0$. These points are also crucial for the topological structure of the MPVC (see the cell-attachment theorem). For global optimization, points of T-index 1 play an important role. We emphasize that among the points of T-index 1 from a topological point of view there is no substantial difference between the points with $BI = 1, QI = 0$ and $BI = 0, QI = 1$. It is worth mentioning that a linear descent direction might exist in a nondegenerate T-stationary point with positive T-index. In particular, at points with $BI = 1, QI = 0$ there are exactly two directions of linear descent. Both of them are important from a global point of view. On the other hand, among weakly stationary points, there are those with negative and positive Lagrange multipliers corresponding to the same bi-active vanishing constraint. Due to the deformation theorem, such points are irrelevant for the topological structure of the MPVC.

We mention that the nondegeneracy assumption (as in Definition 30, ND4) cannot be stated for M- and strongly stationary points w.r.t. biactive vanishing constraints. This means that these points are singularities. Moreover, local minima for MPVC with bi-active vanishing constraints do not occur generically. We claim that their classification is sophisticated and might be established via singularity theory.

Links to MPCC

We point out that in Section 2.3 (see also [69]) the analogous stationarity concept for MPCCs turned out to be C-stationarity. Indeed, the MPCC feasible set can be described by nonsmooth equality constraints of minimum type. Moreover, generically the MPCC feasible set is a Lipschitz manifold of an appropriate dimension; that is, each nonsmooth equality constraint causes loss of one degree of freedom (see Section 2.2.2 and [70]). This permits the use of Clarke subdifferentials of these equality constraints to formulate the stationarity conditions, namely the C-stationarity. As C-stationarity is the topologically relevant stationarity concept for MPCCs, we consider it T-stationarity in the MPCC setting.

In contrast to the MPCC case, the MPVC feasible set (under the LICQ) is not a Lipschitz manifold but a set glued together from manifold pieces of **different dimensions** along their strata. Rather than by applying a general stationarity concept to MPVCs, like C-stationarity for MPCCs, T-stationarity for MPVCs is motivated by understanding the geometrical properties of a typical MPVC feasible set $\mathbb{V}$ directly, where $\mathbb{V}$ represents the solution set of the basic vanishing constraint relations $x \geq 0, xy \geq 0$.

A further analogy between C-stationarity for MPCCs and T-stationarity for MPVCs is established via convergence theory of certain regularization methods. In fact, the MPCC regularization method from [108] yields C-stationary points as limits of KKT points of the regularized problems ([108, Theorem 5.1]). The analogous limit points of an adaptation of this method to MPVCs from [47] are T-stationary.