Deutsche Forschungsgemeinschaft
Priority Programme 1962

On Error Bounds and Multiplier Methods for Variational Problems in Banach Spaces

Christian Kanzow, Daniel Steck

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

Preprint Number SPP1962-034

received on September 6, 2017

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

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

On Error Bounds and Multiplier Methods for Variational Problems in Banach Spaces*

Christian Kanzow† Daniel Steck†

September 6, 2017

Abstract. This paper deals with a general form of variational problems in Banach spaces which encompasses variational inequalities as well as minimization problems. We prove a characterization of local error bounds for the distance to the (primal-dual) solution set and give a sufficient condition for such an error bound to hold. In the second part of the paper, we consider an algorithm of augmented Lagrangian type for the solution of such variational problems. We give some global convergence properties of the method and then use the error bound theory to provide estimates for the rate of convergence and to deduce boundedness of the sequence of penalty parameters. Finally, we present some numerical results for optimal control and Nash equilibrium problems.

Keywords. Variational problem, error bound, augmented Lagrangian method, local convergence, global convergence.

1 Introduction

This paper deals with the following variational inequality problem (VI):

$$\text{Find } x \in M \text{ such that } \quad \langle F(x), v \rangle \geq 0 \quad \forall v \in \mathcal{T}_M(x), \tag{1}$$

where $M \subseteq X$ is a nonempty closed set, X a real Banach space, and $F : X \to X^*$ a given mapping. The set $\mathcal{T}_M(x)$ denotes the (Bouligand) tangent cone [4, 9] to M at x. If M is additionally convex, then (1) is equivalent to

$$\text{Find } x \in M \text{ such that } \quad \langle F(x), y - x \rangle \geq 0 \quad \forall y \in M, \tag{2}$$

which is often regarded as the standard form of a variational inequality. However, in the absence of convexity, (1) is the canonical formulation of variational problems; in particular,

*This research was supported by the German Research Foundation (DFG) within the priority program "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization" (SPP 1962) under grant number KA 1296/24-1.

†University of Würzburg, Institute of Mathematics, Campus Hubland Nord, Emil-Fischer-Str. 30, 97074 Würzburg, Germany; {kanzow,daniel.steck}@mathematik.uni-wuerzburg.de

this general form encompasses first-order necessary conditions for nonlinear optimization problems of the type

$$\min\ f(x) \quad \text{s.t.} \quad x \in M \tag{3}$$

by choosing $F := f'$. Throughout this paper, we assume that M is given in the form

$$M = \{x \in X : g(x) \in K\}, \tag{4}$$

where $g : X \to H$ is a given mapping, H a real Hilbert space, and $K \subseteq H$ a nonempty closed convex set (not necessarily a cone). We make no blanket convexity assumptions on g (although some of our results do pertain to the convex case). Hence, the set M is nonconvex in general, and we are forced to deal with the VI in the form (1).

Variational inequalities are a well-known and popular class in both finite and infinite-dimensional optimization since they unify various problem types such as constrained minimization and equilibrium-type problems, in particular Nash and (certain) generalized Nash equilibrium problems [15, 16, 20, 28, 35]. Many further applications are given in [3, 24, 25, 38]. As a result, VIs have gained considerable attention in the literature and a variety of algorithms have been developed for their solution, e.g. [17, 22, 23, 47].

On the other hand, the augmented Lagrangian method (ALM, also called multiplier-penalty method or simply multiplier method) is one of the classical methods for nonlinear optimization, see [11, 26, 43, 44, 45] and the textbooks [5, 41]. In recent years, ALMs have seen a certain resurgence [1, 2, 6, 7, 8] in the form of modified methods which use a slightly different update of the Lagrange multiplier and turn out to have very strong global convergence properties [8]. A comparison of the classical and modified ALMs is given in [36]. We also note that ALMs have been generalized to VIs in finite dimensions [2] and to infinite-dimensional optimization problems in certain restricted settings [27, 30, 31, 32, 33, 37, 49]. However, most of these papers either use very strong assumptions [27], consider very specific problem settings [30, 31, 32, 33], or deal with global convergence properties only [37].

The main purpose of the present paper is to analyze the local convergence properties of ALMs for variational inequalities in the general (possibly infinite-dimensional) setting (1). To accomplish this, we will need certain elements of perturbation and error bound theory for generalized equations and KKT systems, some of which are refinements of the corresponding results in finite dimensions [12, 13, 19, 34]. Using these, we will prove that, given a KKT point which admits a primal-dual error bound, the ALM converges locally to this point with a rate of convergence that is essentially $1/\rho_k$ (where ρ_k is the penalty parameter), and that $\{\rho_k\}$ remains bounded if updated suitably.

Sufficient conditions for the primal-dual error bound include a suitable second-order sufficient condition (SOSC) together with a strict version of the Robinson constraint qualification (see Section 2). These assumptions are akin to those used in [6] for ALMs in finite-dimensional nonlinear programming (NLP), where the authors obtain results similar to ours. Interestingly, however, it turns out that these results can be established under SOSC only [18] by using the specific structure of NLP constraints. In particular, when transferred to our notation, the set K arising from NLP is polyhedral and this yields, roughly speaking, the dual part of the error bound without any constraint qualification [18, 34]. However, apart from the NLP setting, polyhedrality is a rare property which is usually violated, e.g. in semidefinite programming or optimal control. As a result, it is quite clear that SOSC alone is not enough for general constraints of the form (4), see also the example in Section 3. We solve this issue by using SOSC together with a suitable constraint qualification.

This paper is organized as follows. We start with some preliminary material in Section 2 and give some results on primal-dual error bounds in Section 3. Section 4 contains a precise statement of our algorithm and we continue with some global convergence results in Section 5. In Section 6, we prove the main results of this paper, i.e. local convergence of the ALM under the error bound hypothesis. We then give some numerical results in Section 7 and final remarks in Section 8.

Notation: Throughout the paper, X is always a real Banach space, H a real Hilbert space, and their duals are denoted by X^* and H^*, the latter of which we usually identify with H. Strong and weak convergence are denoted by $\to$ and $\rightharpoonup$, respectively. Duality pairings are written as $\langle \cdot, \cdot \rangle$, scalar products as $(\cdot, \cdot)$, and norms are denoted by $\| \cdot \|$ with an appropriate subscript to emphasize the corresponding space (e.g. $\| \cdot \|_X$). If S is a nonempty subset of some normed space, we write $d_S = \operatorname{dist}(\cdot, S)$ for the distance to S. Additionally, if $S \subseteq H$ is closed and convex, we write P_S for the projection onto S.

2 Preliminaries

This section is dedicated to establishing some preliminary results as well as fixing the setting we will consider later. Recall that the set M is given by the formula (4) with a nonempty closed convex set $K \subseteq H$.

2.1 Cones and Convexity

If S is a nonempty closed subset of some space Z, then $S^\circ := \{\psi \in Z^* : \langle \psi, s \rangle \le 0 \ \forall s \in S\}$ denotes the *polar cone* of S. If Z is a Hilbert space, we of course treat S° as a subset of Z. Moreover, if $x \in S$ is a given point, we denote by

$$\mathcal{T}_S(x) := \left\{ \lim_{k\to\infty} \frac{x^k - x}{t_k} : \{x^k\} \subseteq S, \ x^k \to x, \ t_k \downarrow 0 \right\}$$

the *tangent cone* of S at x. If S is additionally convex, we also define the *normal cone*

$$\mathcal{N}_S(x) := \{\psi \in Z^* : \langle \psi, y - x \rangle \le 0 \ \forall y \in S\} = \mathcal{T}_S(x)^\circ.$$

If $x \notin S$, we define $\mathcal{T}_S(x)$ and $\mathcal{N}_S(x)$ to be empty. Note that, if S is a convex set, then $\mathcal{T}_S(x)$ and $\mathcal{N}_S(x)$ are closed convex cones for all $x \in S$.

Recall that the constraint system of the VI is given by $g(x) \in K$ with $K \subseteq H$ a nonempty closed convex set. A natural question is what the appropriate notion of convexity is in this general setting. In particular, we would like to give sufficient conditions for the convexity of the feasible set M. To this end, consider the recession cone

$$K_\infty := \{y \in H : y + K \subseteq K\}. \tag{5}$$

It is well-known that K_∞ is a nonempty closed convex cone [4, 9]. If K itself is a cone, then $K_\infty = K$. We associate with K (and K_∞) the order relation

$$y \le_K z :\Longleftrightarrow z - y \in K_\infty. \tag{6}$$

Note that we use the notation $\leq_K$ for the sake of convenience, even though the order is actually induced by the cone K_∞. We also note that K_∞ may not be pointed (that is, $K_\infty \cap (-K_\infty)$ may be nonempty) and, hence, the relation $\leq_K$ does not necessarily satisfy the antisymmetry property

$$a \leq_K b \wedge b \leq_K a \implies a = b.$$

In the terminology of order theory, this makes $\leq_K$ a so-called preorder. We will simply call it an order relation due to the descriptiveness of the term. Note also that, throughout this paper, the symbol $\leq$ without any index is always the standard ordering in $\mathbb{R}$.

The order relation (6) allows us to extend various familiar concepts from finite-dimensional optimization to our setting. For instance, we say that g is convex if

$$g(\alpha x + (1-\alpha)y) \leq_K \alpha g(x) + (1-\alpha)g(y)$$

holds for all $x, y \in X$ and $\alpha \in [0,1]$. Other notions which involve an order such as increasing, decreasing or concave functions are also defined in a straightforward way. For example, the distance function $d_K : H \to \mathbb{R}$ is decreasing since $z \geq_K y$ implies $z = y + k$, $k \in K_\infty$, and

$$d_K(z) = d_K(y+k) \leq \|y + k - (P_K(y) + k)\| = \|y - P_K(y)\| = d_K(y),$$

where the inequality uses the fact that $P_K(y) + k \in K$ by definition of K_∞. Some other results pertaining to convexity, concavity, etc. are given in the following lemma. Note that, in the context of our constraint set (4) with $g(x) \in K$, it is more natural to consider concavity of g with respect to the ordering (6) as opposed to convexity.

Lemma 2.1. *Assume that $g : X \to H$ is concave. If $m : H \to \mathbb{R}$ is convex and decreasing, then $m \circ g$ is convex. In particular:*

(a) *The function $d_K \circ g : X \to \mathbb{R}$ is convex.*

(b) *If $\lambda \in K_\infty^\circ$, then $x \mapsto (\lambda, g(x))$ is convex.*

(c) *The set $M = \{x \in X : g(x) \in K\}$ is convex.*

Proof. Let $x, y \in X$ and $x_\alpha = \alpha x + (1-\alpha)y$, $\alpha \in [0,1]$. Then $g(x_\alpha) \geq_K \alpha g(x) + (1-\alpha)g(y)$ by the concavity of g. Applying m on both sides yields

$$m(g(x_\alpha)) \leq m(\alpha g(x) + (1-\alpha)g(y)) \leq \alpha m(g(x)) + (1-\alpha)m(g(y)),$$

where we used the monotonicity and the convexity of m. Hence, $m \circ g$ is convex. Assertion (a) now follows because d_K is decreasing (see above) and convex [4, Cor. 12.12]. Similarly, for (b), the function $y \mapsto (\lambda, y)$ with $\lambda \in K_\infty^\circ$ is obviously a convex function, and it is decreasing because $(\lambda, k) \leq 0$ for all $k \in K_\infty$. Finally, for (c), note that

$$M = \{x \in X : g(x) \in K\} = \{x \in X : d_K(g(x)) \leq 0\}.$$

Hence, M is a lower level set of the convex function $d_K \circ g$ and therefore a convex set. $\square$

Note that the extreme case $K_\infty = \{0\}$ can occur, e.g. if K is bounded. In this case, monotonicity becomes trivial and convexity and concavity reduce to linearity.

It is possible to characterize K_∞° by means of the so-called barrier cone to K, see [4]. Here, we will only need the following observation.

Lemma 2.2. *If $y \in H$, then $y - P_K(y) \in K_\infty^\circ$.*

Proof. Let $k \in K$ be fixed and let $z \in K_\infty$, $\alpha \geq 0$. Then $\alpha z \in K_\infty$ and therefore $\alpha z + k \in K$. A standard projection inequality yields $(y - P_K(y), \alpha z + k - P_K(y)) \leq 0$. But this clearly cannot hold for all α if $(y - P_K(y), z) > 0$. Hence, $(y - P_K(y), z) \leq 0$. $\square$

2.2 Optimality Conditions

We now turn to the variational inequality (1) and discuss its KKT conditions. Starting with this section, we assume that the mapping F is continuously differentiable and that g is twice continuously differentiable. Consider now the Lagrange function

$$\mathcal{L} : X \times H \to X^*, \quad \mathcal{L}(x, \lambda) := F(x) + g'(x)^* \lambda. \tag{7}$$

Note that, if the VI originates from a minimization problem, then $\mathcal{L}$ is actually the derivative of the conventional Lagrange function. The following are the standard first-order optimality conditions which we will use throughout this paper.

Definition 2.3. A tuple $(\bar{x}, \bar{\lambda}) \in X \times H$ is a *KKT point* of (1), (4) if

$$\mathcal{L}(\bar{x}, \bar{\lambda}) = 0 \quad \text{and} \quad \bar{\lambda} \in \mathcal{N}_K(g(\bar{x})). \tag{8}$$

We call $\bar{x} \in X$ a stationary point if $(\bar{x}, \bar{\lambda})$ is a KKT point for some $\bar{\lambda} \in H$, and denote by $\mathcal{M}(\bar{x})$ the corresponding set of multipliers.

Note that $\bar{\lambda} \in \mathcal{N}_K(g(\bar{x}))$ implies $g(\bar{x}) \in K$, since otherwise the normal cone would be empty. Moreover, we remark that, if K is a cone, then $\bar{\lambda} \in \mathcal{N}_K(g(\bar{x}))$ is equivalent to the complementarity conditions $g(\bar{x}) \in K$, $\bar{\lambda} \in K^\circ$, and $\big(\bar{\lambda}, g(\bar{x})\big) = 0$, see [9, Ex. 2.62].

The relationship between the VI and its KKT conditions is given as follows: if $\bar{x}$ solves the VI and a suitable constraint qualification holds in $\bar{x}$, then there exists a multiplier $\bar{\lambda}$ such that $(\bar{x}, \bar{\lambda})$ is a KKT point [9, Remark 5.8]. On the other hand, it is easy to see that the KKT conditions are always sufficient optimality conditions, even if M is nonconvex. This result is contained in the following theorem and crucially depends on the fact that the VI uses the tangent cone $\mathcal{T}_M$ and not M itself.

Theorem 2.4. *Every KKT point $(\bar{x}, \bar{\lambda})$ of the VI is a solution of the VI.*

Proof. Let $(\bar{x}, \bar{\lambda})$ be a KKT point and $d \in \mathcal{T}_M(\bar{x})$. Then $d = \lim_{k\to\infty}(x^k - \bar{x})/t_k$ with $\{x^k\} \subseteq M$, $x^k \to \bar{x}$, and $t_k \downarrow 0$. Hence,

$$\langle F(\bar{x}), d\rangle = \left\langle -g'(\bar{x})^*\bar{\lambda}, \lim_{k\to\infty} \frac{x^k - \bar{x}}{t_k} \right\rangle = -\lim_{k\to\infty} \frac{1}{t_k}\big(\bar{\lambda}, g'(\bar{x})(x^k - \bar{x})\big).$$

But $g'(\bar{x})(x^k - \bar{x}) = g(x^k) - g(\bar{x}) + o(t_k)$ and therefore

$$\langle F(\bar{x}), d\rangle = -\lim_{k\to\infty} \frac{1}{t_k}\big(\bar{\lambda}, g(x^k) - g(\bar{x})\big) \geq 0,$$

where we used $\bar{\lambda} \in \mathcal{N}_K(g(\bar{x}))$ and $g(x^k) \in K$ for all k. $\square$

For a given KKT point $(\bar{x}, \bar{\lambda})$ and $\eta \geq 0$, we define the extended critical cone

$$C_\eta(\bar{x}) := \big\{ d \in X : \langle F(\bar{x}), d \rangle \leq \eta \|d\|_X,\ g'(\bar{x})d \in \mathcal{T}_K(g(\bar{x})) \big\}.$$

The following is the second-order condition which we will use throughout this paper.

Definition 2.5. Let $(\bar{x}, \bar{\lambda})$ be a KKT point of the VI. We say that the *second-order sufficient condition (SOSC)* holds in $(\bar{x}, \bar{\lambda})$ if there are $\eta, c > 0$ such that

$$\big\langle \mathcal{L}'(\bar{x}, \bar{\lambda})d, d \big\rangle \geq c\|d\|_X^2 \quad \text{for all } d \in C_\eta(\bar{x}).$$

Note that we use the terminology "second-order sufficient condition" mainly for the sake of consistency with a similar condition from nonlinear optimization, e.g. [9, Def. 3.60]. For variational problems such as (1), there is actually no need for sufficiency conditions to complement the KKT system because the latter is always sufficient for optimality by Theorem 2.4.

Let us also note that Definition 2.5 is slightly different from the second-order sufficient condition for nonlinear optimization [9, Def. 3.60] because our extended critical cone is slightly smaller. However, under the Robinson constraint qualification (see below and [9, Def. 2.86]), the corresponding second-order conditions coincide [9, Remark 3.68]. Moreover, and more importantly, our subsequent analysis will be based on [9, Thm. 5.9] which directly uses the "smaller" critical cone together with the following constraint qualification.

Definition 2.6. Let $(\bar{x}, \bar{\lambda})$ be a KKT point of the VI, and let $K_0 := \{y \in K : \big(\bar{\lambda}, y - g(\bar{x})\big) = 0\}$. We say that the *strict Robinson condition (SRC)* holds in $(\bar{x}, \bar{\lambda})$ if

$$0 \in \operatorname{int}\big(g(\bar{x}) + g'(\bar{x})X - K_0\big). \tag{9}$$

Note that the standard Robinson constraint qualification arises if we replace K_0 in (9) by the larger set K. Hence, SRC is stronger than the Robinson constraint qualification and the equivalent regularity condition of Zowe and Kurcyusz [50]. On the other hand, SRC implies the uniqueness of $\bar{\lambda}$ and is weaker than the surjectivity of $g'(\bar{x})$, which is a typical regularity assumption for infinite-dimensional problems.

It should be noted that the definition of SRC presupposes the existence of $\bar{\lambda}$ and therefore depends not only on the constraints but also on the function F. Hence, we refrain from calling (9) a constraint qualification (in contrast to [9], where SRC is called the *strict constraint qualification*). A similar condition which is occasionally used in the finite-dimensional literature is the strict Mangasarian-Fromovitz condition [6, 21, 39]. This condition turns out to be a special case of SRC [9, Remark 4.49] and is also not a constraint qualification [48].

3 Error Bounds for the Variational Problem

Recall that the KKT conditions of the VI are given by

$$\mathcal{L}(\bar{x}, \bar{\lambda}) = 0 \quad \text{and} \quad \bar{\lambda} \in \mathcal{N}_K(g(\bar{x})),$$

where $(\bar{x}, \bar{\lambda}) \in X \times H$. The last condition is well-known [4, Prop. 6.46] to be equivalent to $g(\bar{x}) = P_K(g(\bar{x}) + \bar{\lambda})$. This suggests defining the residual mapping

$$\sigma(x, \lambda) := \|\mathcal{L}(x, \lambda)\|_{X^*} + \|g(x) - P_K(g(x) + \lambda)\|_H. \tag{10}$$

Clearly, the KKT conditions of the VI are equivalent to $\sigma(\bar{x}, \bar{\lambda}) = 0$. We will use this relationship to construct suitable error bounds for the primal-dual variables.

In order to establish the error bound we are looking for, we first need a characterization of local error bounds in terms of a local upper Lipschitz property (also called isolated calmness) of the KKT system. This result has appeared in various forms in the literature [12, 19, 34], albeit mostly in a finite-dimensional setting. In our notation, it involves certain perturbations of the KKT system (8) with a parameter pair $\sigma = (\alpha, \beta) \in X^* \times H$. Without loss of generality, we equip this product space with the norm $\|(\alpha, \beta)\|_{X^* \times H} := \|\alpha\|_{X^*} + \|\beta\|_H$. Recall also that $\mathcal{M}(\bar{x})$ denotes the set of Lagrange multipliers corresponding to $\bar{x}$.

Theorem 3.1. *Let $(\bar{x}, \bar{\lambda}) \in X \times H$ be a KKT point of the VI. Then the following assertions are equivalent:*

(a) *There are a neighborhood U of $\bar{x}$ and $c > 0$ such that, for all $\sigma = (\alpha, \beta) \in X^* \times H$ close to $(0, 0)$, any solution $(x_\sigma, \lambda_\sigma) \in U \times H$ of the perturbed KKT system*

$$\mathcal{L}(x, \lambda) = \alpha, \quad \lambda \in \mathcal{N}_K(g(x) - \beta) \tag{11}$$

satisfies the estimate $\|x_\sigma - \bar{x}\|_X + \operatorname{dist}(\lambda_\sigma, \mathcal{M}(\bar{x})) \le c\|\sigma\|_{X^* \times H}$.

(b) *There are a neighborhood U of $\bar{x}$ and $c > 0$ such that, for all $(x, \lambda) \in U \times H$ with $\sigma(x, \lambda)$ sufficiently small,*

$$\|x - \bar{x}\|_X + \operatorname{dist}(\lambda, \mathcal{M}(\bar{x})) \le c\sigma(x, \lambda).$$

Proof. (b)⇒(a): Let $\sigma = (\alpha, \beta) \in X^* \times H$. It is an easy consequence of [4, Cor. 4.10] that the mapping $y \mapsto y - P_K(y + \lambda_\sigma)$ is nonexpansive. Hence, we obtain the inequality

$$\|g(x_\sigma) - P_K(g(x_\sigma) + \lambda_\sigma)\|_H \le \|\beta\|_H + \|g(x_\sigma) - \beta - P_K(g(x_\sigma) - \beta + \lambda_\sigma)\|_H.$$

Since $\lambda_\sigma \in \mathcal{N}_K(g(x_\sigma) - \beta)$, the last term is equal to zero [4, Prop. 6.46] and we obtain $\sigma(x_\sigma, \lambda_\sigma) \le \|\alpha\|_{X^*} + \|\beta\|_H = \|\sigma\|_{X^* \times H}$. Choosing $\sigma = (\alpha, \beta)$ sufficiently close to 0, we see that $\sigma(x_\sigma, \lambda_\sigma)$ becomes arbitrarily small. Hence, we can apply (b) and obtain

$$\|x_\sigma - \bar{x}\|_X + \operatorname{dist}(\lambda_\sigma, \mathcal{M}(\bar{x})) \le c\sigma(x_\sigma, \lambda_\sigma) \le c\|\sigma\|_{X^* \times H}.$$

(a)⇒(b): Shrinking U if necessary, we may assume that $\|g'(x)^*\|_{\mathcal{L}(H, X^*)} \le c_1$ for all $x \in U$ with some constant $c_1 \ge 0$. Let $(x, \lambda) \in U \times H$, set $\delta := \sigma(x, \lambda)$, and define

$$\hat{g} := P_K(g(x) + \lambda), \quad \hat{\lambda} := g(x) + \lambda - \hat{g}.$$

Now, let $\alpha := \mathcal{L}(x, \hat{\lambda})$ and $\beta := g(x) - \hat{g}$. Then $\hat{\lambda} \in \mathcal{N}_K(\hat{g})$ and, hence, $(x, \hat{\lambda})$ solves the perturbed KKT system corresponding to $\sigma := (\alpha, \beta)$. Moreover, we have $\|\beta\|_H = \|\hat{g} - g(x)\|_H = \|g(x) - P_K(g(x) + \lambda)\|_H \le \delta$ and $\|\hat{\lambda} - \lambda\|_H = \|\beta\|_H \le \delta$. This implies

$$\|\sigma\|_{X^* \times H} = \|\mathcal{L}(x, \hat{\lambda})\|_{X^*} + \|\beta\|_H \le \|\mathcal{L}(x, \lambda)\|_{X^*} + (c_1 + 1)\|\beta\|_H \le (c_1 + 2)\delta.$$

Hence, if $\delta = \sigma(x, \lambda)$ is small enough, then σ becomes arbitrarily close to 0. We can therefore apply (a) to $(x, \hat{\lambda})$ and obtain

$$\|x - \bar{x}\|_X + \operatorname{dist}(\hat{\lambda}, \mathcal{M}(\bar{x})) \le c\|\sigma\|_{X^* \times H} \le c(c_1 + 2)\delta.$$

But $\|\hat{\lambda} - \lambda\|_H \le \delta$ and, hence, $\operatorname{dist}(\hat{\lambda}, \mathcal{M}(\bar{x})) \ge \operatorname{dist}(\lambda, \mathcal{M}(\bar{x})) - \delta$ by the nonexpansiveness of the distance function. This finally yields

$$\|x - \bar{x}\|_X + \operatorname{dist}(\lambda, \mathcal{M}(\bar{x})) \le \big[c(c_1 + 2) + 1\big]\delta,$$

and the proof is complete. □

Let us stress that the distance estimate provided by the above theorem holds if x is close to $\bar{x}$; in particular, no assumption on the proximity of λ to $\mathcal{M}(\bar{x})$ is necessary. We also remark that (a) does not make any assertion about the existence of solutions to the perturbed KKT conditions (11). These may have solutions for some but not all σ.

Theorem 3.1 is our main tool for establishing local error bounds for the distance of (x, λ) to the primal-dual solution set in terms of the residual mapping σ. To verify such an error bound, we only need to prove property (a) of the theorem. The following result does precisely that and is based on the perturbation theory from [9].

Theorem 3.2. *Assume that $(\bar{x}, \bar{\lambda})$ is a KKT point which satisfies SOSC and the strict Robinson condition. Then $\mathcal{M}(\bar{x}) = \{\bar{\lambda}\}$ and there is a $c > 0$ such that, for all $(x, \lambda) \in X \times H$ with x sufficiently close to $\bar{x}$,*

$$\|x - \bar{x}\|_X + \|\lambda - \bar{\lambda}\|_H \le c\sigma(x, \lambda). \tag{12}$$

Proof. The uniqueness of $\bar{\lambda}$ follows as in [9, Prop. 4.47], see also the discussion in Section 5.1.2 of that reference. For the error bound result, we essentially need to apply [9, Thm. 5.9] and Theorem 3.1. Since some technical details need to be considered, we give a formal proof here. To this end, assume that the error bound in question does not hold. Then property (a) from Theorem 3.1 does not hold either; hence, there are sequences $x^k \to \bar{x}$, $\{\lambda^k\} \subseteq H$ and $\{\sigma^k\} \subseteq X^* \times H$ with $\sigma^k = (\alpha^k, \beta^k) \to 0$ such that, for all k, (x^k, λ^k) satisfies the perturbed KKT conditions (11) corresponding to σ^k, and

$$\|x^k - \bar{x}\|_X + \|\lambda^k - \bar{\lambda}\|_H \ge k\|\sigma^k\|_{X^* \times H}. \tag{13}$$

Now, let $F(x, \sigma) := F(x) - \alpha$ and $G(x, \sigma) := g(x) - \beta$ for $\sigma = (\alpha, \beta) \in X^* \times H$. Then (x^k, λ^k) satisfies

$$F(x^k, \sigma^k) + D_x G(x^k, \sigma^k)^* \lambda^k = 0, \quad \lambda^k \in \mathcal{N}_K(G(x^k, \sigma^k))$$

for all k. Applying [9, Thm. 5.9] yields a contradiction to (13). □

The function σ is locally Lipschitz-continuous with respect to (x, λ), and globally so with respect to λ. Hence, we can extend the one-sided error bound (12) to

$$c_1\sigma(x, \lambda) \le \|x - \bar{x}\|_X + \|\lambda - \bar{\lambda}\|_H \le c_2\sigma(x, \lambda) \tag{14}$$

for suitable constants $c_1, c_2 > 0$ and all $(x, \lambda) \in X \times H$ with x near $\bar{x}$.

For certain problem classes, it is possible to establish error bounds under weaker assumptions than those given above. The most important example in this direction is if the set K is (generalized) polyhedral, e.g. in nonlinear programming. Roughly speaking, one can use Hoffman's lemma [9, Thm. 2.200] to get the "dual part" of the error bound for free, while the primal part again follows from SOSC. As a result, one obtains a primal-dual error bound under SOSC alone (with the restriction that the multiplier is not necessarily unique). Unsurprisingly, this result does not extend to the non-polyhedral case, which shows that additional assumptions such as SRC are inevitable.

Example 3.3. Let $X := H := \ell^2(\mathbb{R})$ be the space of square-summable real sequences. Consider the optimization problem (3), (4) with $f(x) := \|x\|_X^2/2$, $g(x) := (x_i/i)_{i=1}^{\infty}$, and K the nonnegative cone in X. It is easy to see that $(\bar{x}, \bar{\lambda}) := (0, 0)$ is the unique KKT point of this problem, and that SOSC holds. Now, let $x^k := e^k/k$ and $\lambda^k := -e^k$, where $\{e^k\}$ is the sequence of unit vectors. Then

$$\sigma(x^k, \lambda^k) = \|\mathcal{L}(x^k, \lambda^k)\|_{X^*} + \|g(x^k) - P_K(g(x^k) + \lambda^k)\|_H = k^{-2}$$

for all k. Moreover, $x^k \to \bar{x}$, but $\lambda^k \not\to \bar{\lambda}$. Hence, a local error bound does not hold. (In particular, SRC cannot hold, even though the Lagrange multiplier is actually unique.) A slightly different example is obtained by setting $\hat{x}^k := e^k/k^2$ and $\hat{\lambda}^k := -e^k/k$. In this case, $(\hat{x}^k, \hat{\lambda}^k) \to (\bar{x}, \bar{\lambda})$, but an easy calculation shows that

$$\sigma(\hat{x}^k, \hat{\lambda}^k) = k^{-3} \quad \text{and} \quad \|\hat{x}^k - \bar{x}\|_X + \|\hat{\lambda}^k - \bar{\lambda}\|_H = k^{-2} + k^{-1}.$$

In particular, the error bound is violated even if the multiplier is close to $\bar{\lambda}$.

4 The Augmented Lagrangian Method

We now present the augmented Lagrangian method for the variational inequality (1). The main approach is to penalize the function g and therefore reduce the VI to a sequence of (unconstrained) nonlinear equations. Consider the augmented Lagrangian

$$\mathcal{L}_\rho : X \times H \to X^*, \quad \mathcal{L}_\rho(x, \lambda) := F(x) + \rho g'(x)^* \left[g(x) + \frac{\lambda}{\rho} - P_K\left(g(x) + \frac{\lambda}{\rho} \right) \right]. \tag{15}$$

Note that, if K is a cone, then we can simplify the above formula to $\mathcal{L}_\rho(x, \lambda) = F(x) + g'(x)^* P_{K^\circ}(\lambda + \rho g(x))$ by using Moreau's decomposition [4, 40].

For the construction of our algorithm, we will need a means of controlling the penalty parameters. To this end, we define the utility function

$$V(x, \lambda, \rho) := \|\mathcal{L}_\rho(x, \lambda)\|_{X^*} + \left\| g(x) - P_K\left(g(x) + \frac{\lambda}{\rho} \right) \right\|_H. \tag{16}$$

This function requires some elaboration. The first term in (16) measures the precision with which the subproblem was solved in the current iteration. The second term is a composite measure of feasibility and complementarity; it arises from an inherent slack variable transformation which is often used to define the augmented Lagrangian for inequality or cone constraints. As a result, the function V measures optimality, feasibility and complementarity at the current iterate.

Algorithm 4.1 (Augmented Lagrangian method)**.**

(S.0) *Let* $(x^0, \lambda^0) \in X \times H$*,* $B \subseteq H$ *bounded,* $\rho_0 > 0$*,* $\gamma > 1$*,* $\tau \in (0, 1)$*, and set* $k := 0$*.*

(S.1) *If* (x^k, λ^k) *satisfies a suitable termination criterion: STOP.*

(S.2) *Choose* $w^k \in B$ *and compute an inexact zero (see below)* x^{k+1} *of* $\mathcal{L}_{\rho_k}(\cdot, w^k)$*.*

(S.3) *Update the vector of multipliers to*

$$\lambda^{k+1} := \rho_k \left[g(x^{k+1}) + \frac{w^k}{\rho_k} - P_K\left(g(x^{k+1}) + \frac{w^k}{\rho_k} \right) \right]. \tag{17}$$

(S.4) *If* $k = 0$ *or*

$$V(x^{k+1}, w^k, \rho_k) \leq \tau V(x^k, w^{k-1}, \rho_{k-1}) \tag{18}$$

holds, set $\rho_{k+1} := \rho_k$*; otherwise, set* $\rho_{k+1} := \gamma \rho_k$*.*

(S.5) *Set* $k \leftarrow k + 1$ *and go to (S.1).*

Let us make some simple observations. First, regardless of the primal iterates $\{x^k\}$, the multipliers $\{\lambda^k\}$ always lie in the polar cone K_∞° by Lemma 2.2. Moreover, if K is a cone, then the Moreau decomposition [40] implies that $\lambda^{k+1} = P_{K^\circ}(w^k + \rho_k g(x^{k+1}))$.

Secondly, we note that Algorithm 4.1 uses a safeguarded multiplier sequence $\{w^k\}$ in certain places where classical augmented Lagrangian methods use the sequence $\{\lambda^k\}$. This bounding scheme goes back to [1, 42] and is crucial to establishing strong global convergence results for the method [1, 7, 8, 37]. In practice, one usually tries to keep w^k as "close" as possible to λ^k, e.g. by defining $w^k := P_B(\lambda^k)$, where B (the bounded set from the algorithm) is chosen suitably to allow cheap projections.

The third observation is that if the sequence of penalty parameters $\{\rho_k\}$ remains bounded, then (18) yields $V(x^{k+1}, w^k, \rho_k) \to 0$. In this case, the definition of V implies that both the residual $\|\mathcal{L}_{\rho_k}(x^{k+1}, w^k)\|_{X^*}$ of the subproblems and the composite feasibility-complementarity measure converge to zero. Hence, from a theoretical point of view, the case of bounded $\{\rho_k\}$ is the "good" case. In Section 6, we will actually prove the boundedness of $\{\rho_k\}$ under certain assumptions, and this result crucially depends on the fact that the function V involves both terms from (16).

For the remainder of this paper, we make the following assumption.

Assumption 4.2. There is a zero sequence $\{\varepsilon_k\} \subseteq [0, \infty)$ such that

$$\|\mathcal{L}_{\rho_k}(x^{k+1}, w^k)\|_{X^*} \leq \varepsilon_{k+1} \quad \text{for all } k.$$

This assumption is fairly natural and basically asserts that x^{k+1} is an approximate zero point of $\mathcal{L}_{\rho_k}(\cdot, w^k)$, and that the degree of inexactness vanishes as $k \to \infty$.

5 Global Convergence

In this section, we discuss the global convergence properties of Algorithm 4.1. Some general results in this direction were obtained in [35, 37] for optimization and generalized Nash equilibrium problems by assuming that the sequence $\{x^k\}$ has a limit point which satisfies a suitable constraint qualification.

Here, we pursue a slightly different approach. Since the constraints occurring in VIs are often convex, we can use this convexity to directly show that (weak) limit points are solutions of the VI. This idea has the advantage that we do not need any constraint qualification (in return, we do not get much information on the sequence $\{\lambda^k\}$).

Recall that we have already assumed F to be continuously differentiable and g twice continuously differentiable (for this section, one degree less would actually be sufficient). We now make the following additional assumptions.

Assumption 5.1. We assume that g is concave with respect to K_∞ (see Section 2) and that $\langle F(x), x - y\rangle$ is weakly lsc with respect to x for all $y \in X$.

The first of the above conditions ensures the convexity of the set M, see Lemma 2.1. The second assumption implies, roughly speaking, that weak limit points of a sequence of "approximate solutions" of the VI are exact solutions. Note that this condition has also been used in certain existence results for VIs [29].

Lemma 5.2. *Let Assumptions 4.2, 5.1 hold, and let $\bar{x}$ be a weak limit point of $\{x^k\}$. Then $\bar{x}$ is a minimizer of the convex function $d_K \circ g$. In particular, if the feasible set M is nonempty, then $\bar{x}$ is feasible.*

Proof. Note that the function $d_K \circ g$ is convex by Lemma 2.1 and continuous, hence weakly lower semicontinuous [4, Thm. 9.1]. If $\{\rho_k\}$ remains bounded, then the penalty updating scheme (18) implies

$$d_K(g(x^{k+1})) \le \left\| g(x^{k+1}) - P_K\left(g(x^{k+1}) + \frac{w^k}{\rho_k} \right) \right\|_H \le V(x^{k+1}, w^k, \rho_k) \to 0$$

and therefore $d_K(g(\bar{x})) = 0$. We now assume that $\rho_k \to \infty$ and define the auxiliary functions $h_k(x) = d_K^2(g(x) + w^k/\rho_k)$. Note that h_k is continuously differentiable [4, Cor. 12.30]. Let $x^{k+1} \rightharpoonup_{\mathcal{K}} \bar{x}$ and assume that there is a point $y \in X$ with $d_K(g(y)) < d_K(g(\bar{x}))$. The weak lower semicontinuity of $d_K \circ g$ and the boundedness of $\{w^k\}$ imply

$$\liminf_{k \in \mathcal{K}} h_k(x^{k+1}) = \liminf_{k \in \mathcal{K}} d_K^2(g(x^{k+1})) \ge d_K^2(g(\bar{x}))$$

and $h_k(y) \to d_K^2(g(y))$. Hence, there is a constant $c_1 > 0$ such that $h_k(x^{k+1}) - h_k(y) \ge c_1$ for all $k \in \mathcal{K}$ sufficiently large. Since h_k is convex by Lemma 2.1, it follows that

$$\left\langle h_k'(x^{k+1}), y - x^{k+1} \right\rangle \le h_k(y) - h_k(x^{k+1}) \le -c_1 \tag{19}$$

for all $k \in \mathcal{K}$ sufficiently large. Now, let $\{\varepsilon_k\}$ be the sequence from Assumption 4.2. Using [4, Cor. 12.30] for the derivative of h_k, we obtain

$$\begin{aligned} -\varepsilon_{k+1} \|y - x^{k+1}\|_X &\le \left\langle \mathcal{L}_{\rho_k}(x^{k+1}, w^k), y - x^{k+1} \right\rangle \\ &= \left\langle F(x^{k+1}), y - x^{k+1} \right\rangle + \frac{\rho_k}{2} \left\langle h_k'(x^{k+1}), y - x^{k+1} \right\rangle. \end{aligned}$$

By Assumption 5.1, the function $\langle F(x), x-y\rangle$ is weakly lsc with respect to x. Hence, there is a constant $c_2 \in \mathbb{R}$ such that $\big\langle F(x^{k+1}), y - x^{k+1}\big\rangle \le c_2$ for all $k \in \mathcal{K}$. This together with (19) implies

$$-\varepsilon_{k+1}\|y - x^{k+1}\|_X \le c_2 - \frac{\rho_k c_1}{2} \to -\infty.$$

Since $\{x^{k+1}\}_{\mathcal{K}}$ is bounded and $\varepsilon_k \to 0$, this is a contradiction. $\square$

Note that Lemma 5.2 guarantees that every weak limit point $\bar{x}$ automatically minimizes the constraint violation even if the feasible set M is empty.

We now prove the optimality of limit points. To this end, we first need a technical lemma which essentially asserts some sort of "approximate normality" of λ^{k+1} with respect to K and $g(x^{k+1})$, the latter not necessarily being an element of K. Note that the result does not require any assumptions but directly follows from the definition of λ^{k+1} as well as the updating scheme (18).

Lemma 5.3. *We have* $\limsup_{k\to\infty}\big(\lambda^{k+1}, y - g(x^{k+1})\big) \le 0$ *for all* $y \in K$.

Proof. Let $y \in K$ and define the sequence $s^{k+1} := P_K(g(x^{k+1}) + w^k/\rho_k)$. Then $s^{k+1} \in K$ and it follows from [4, Prop. 6.46] that $\lambda^{k+1} \in \mathcal{N}_K(s^{k+1})$. Moreover, we have

$$g(x^{k+1}) = \frac{\lambda^{k+1} - w^k}{\rho_k} + s^{k+1}. \tag{20}$$

This yields

$$\begin{aligned}\big(\lambda^{k+1}, y - g(x^{k+1})\big) &= \left(\lambda^{k+1}, y - \frac{1}{\rho_k}(\lambda^{k+1} - w^k) - s^{k+1}\right)\\ &\le \frac{1}{\rho_k}\Big[\big(\lambda^{k+1}, w^k\big) - \|\lambda^{k+1}\|_H^2\Big],\end{aligned} \tag{21}$$

where we used $\lambda^{k+1} \in \mathcal{N}_K(s^{k+1})$ for the last inequality. Now, if $\{\rho_k\}$ is bounded, then (18) and (20) imply $\|\lambda^{k+1} - w^k\|_H/\rho_k \to 0$ and therefore $\|\lambda^{k+1} - w^k\|_H \to 0$. This yields the boundedness of $\{\lambda^{k+1}\}$ in H as well as $\big(\lambda^{k+1}, w^k\big) - \|\lambda^{k+1}\|_H^2 = \big(\lambda^{k+1}, w^k - \lambda^{k+1}\big) \to 0$. Hence, the desired results follows from (21). We now assume that $\rho_k \to \infty$. Note that (21) is a quadratic function in λ. A simple calculation therefore shows that

$$\big(\lambda^{k+1}, y - g(x^{k+1})\big) \le \frac{1}{4\rho_k}\|w^k\|_H^2.$$

The boundedness of $\{w^k\}$ now implies $\limsup_{k\to\infty}\big(\lambda^{k+1}, y - g(x^{k+1})\big) \le 0$. $\square$

The above result can be stated more concisely if K is a cone. By inserting $0 \in K$ into the inequality, it is easy to see that it is equivalent to $\liminf_{k\to\infty}\big(\lambda^{k+1}, g(x^{k+1})\big) \ge 0$.

We now turn to the main global convergence result.

Theorem 5.4. *Let Assumptions 4.2, 5.1 hold, and let $\bar{x}$ be a weak limit point of $\{x^k\}$. If the feasible set M is nonempty, then $\bar{x}$ is feasible and solves the VI.*

Proof. Let $x^{k+1} \rightharpoonup_{\mathcal{K}} \bar{x}$ for some $\mathcal{K} \subseteq \mathbb{N}$. The feasibility claim follows from Lemma 5.2. For the optimality, let $y \in M$ be any feasible point. Then $\big\langle \mathcal{L}_{\rho_k}(x^{k+1}, w^k), y - x^{k+1} \big\rangle \geq -\varepsilon_{k+1} \|y - x^{k+1}\|_X$ by Assumption 4.2 and, since $\mathcal{L}_{\rho_k}(x^{k+1}, w^k) = \mathcal{L}(x^{k+1}, \lambda^{k+1})$, we get

$$\begin{aligned} -\varepsilon_{k+1}\|y - x^{k+1}\|_X &\leq \big\langle F(x^{k+1}) + g'(x^{k+1})^* \lambda^{k+1}, y - x^{k+1} \big\rangle \\ &\leq \big\langle F(x^{k+1}), y - x^{k+1} \big\rangle + \big(\lambda^{k+1}, g(y) - g(x^{k+1})\big), \end{aligned}$$

where we used the fact that $x \mapsto \big(\lambda^{k+1}, g(x)\big)$ is convex by Lemma 2.1 (recall that $\lambda^{k+1} \in K_\infty^\circ$). Using $\varepsilon_k \to 0$ and Lemma 5.3, we now obtain $\liminf_{k \in \mathcal{K}} \big\langle F(x^{k+1}), y - x^{k+1} \big\rangle \geq 0$. Since $\langle F(x), x - y \rangle$ is weakly lower semicontinuous, this implies $\langle F(\bar{x}), y - \bar{x} \rangle \geq 0$. □

6 Local Convergence

We will now consider the local convergence characteristics of Algorithm 4.1. A key ingredient is the error bound property from Section 3 which allows us to estimate the distance from (x^k, λ^k) to $(\bar{x}, \bar{\lambda})$ by using the function σ from (10).

Lemma 6.1. *Let Assumption 4.2 hold and let $(\bar{x}, \bar{\lambda})$ be a KKT point satisfying the error bound* (12). *Then there is an $r > 0$ such that, if $x^k \in B_r(\bar{x})$ for all k and $d_K(g(x^k)) \to 0$, then $(x^k, \lambda^k) \to (\bar{x}, \bar{\lambda})$.*

Proof. By Assumption 4.2, we have $\mathcal{L}(x^{k+1}, \lambda^{k+1}) = \mathcal{L}_{\rho_k}(x^{k+1}, w^k) \to 0$. Hence, in view of the error bound property, it suffices to show that $g(x^{k+1}) - P_K(g(x^{k+1}) + \lambda^{k+1}) \to 0$. To this end, define the sequence $s^{k+1} := P_K(g(x^{k+1}) + w^k/\rho_k)$. Then $s^{k+1} \in K$ and, as noted before, $\lambda^{k+1} \in \mathcal{N}_K(s^{k+1})$ for all k. We now use the fact that $y \mapsto y - P_K(y + \lambda^{k+1})$ is nonexpansive, which is an easy consequence of [4, Cor. 4.10]. Therefore, the inverse triangle inequality yields

$$\begin{aligned} &\|g(x^{k+1}) - P_K(g(x^{k+1}) + \lambda^{k+1})\|_H \\ &\leq \|g(x^{k+1}) - s^{k+1}\|_H + \|s^{k+1} - P_K(s^{k+1} + \lambda^{k+1})\|_H. \end{aligned} \tag{22}$$

The last term is equal to zero since $\lambda^{k+1} \in \mathcal{N}_K(s^{k+1})$, cf. [4, Cor. 6.46]. Hence, to complete the proof, we only need to show that $\|s^{k+1} - g(x^{k+1})\|_H \to 0$. If $\{\rho_k\}$ is bounded, then this readily follows from the penalty updating scheme (18). On the other hand, if $\rho_k \to \infty$, then

$$\|s^{k+1} - g(x^{k+1})\|_H \leq \|s^{k+1} - P_K(g(x^{k+1}))\|_H + d_K(g(x^{k+1})) \to 0,$$

where we used the nonexpansiveness of the projection operator. □

The above lemma gives us some information about the behavior of zeros of the augmented Lagrangian in a neighborhood of $\bar{x}$. Note that the assumption $d_K(g(x^{k+1})) \to 0$ asserts that the iterates become (asymptotically) feasible and is often satisfied in practice, see also Lemma 5.2. For the remaining analysis, we now make the following assumption.

Assumption 6.2. We assume that $(\bar{x}, \bar{\lambda})$ is a KKT point of the VI which satisfies the local error bound (12). Moreover, the sequence $\{(x^k, \lambda^k)\}$ from Algorithm 4.1 converges strongly to $(\bar{x}, \bar{\lambda})$, and we have $w^k = \lambda^k$ for all k sufficiently large.

One of the above assumptions which might require some elaboration is $w^k = \lambda^k$ for all k (sufficiently large). The boundedness of $\{w^k\}$ is key to establishing global convergence of the algorithm, see Section 5. Since $\lambda^k \to \bar{\lambda}$ in our setting, we do not need to force boundedness of $\{w^k\}$ and can simply set $w^k := \lambda^k$ for all k. (In the context of Algorithm 4.1, we formally need to choose the bounded set B sufficiently large to allow this.)

We will now prove convergence rates for the primal-dual sequence $\{(x^k, \lambda^k)\}$. Since the distance of (x^k, λ^k) to $(\bar{x}, \bar{\lambda})$ admits both upper and lower estimates relative to the residual terms $\sigma_k := \sigma(x^k, \lambda^k)$ by (14), we will largely base our analysis on the sequence $\{\sigma_k\}$, and the results on the primal-dual sequence $\{(x^k, \lambda^k)\}$ will follow directly.

Lemma 6.3. *Let Assumptions 4.2, 6.2 hold, and let $\sigma_k := \sigma(x^k, \lambda^k)$. Then there is a constant $c_1 > 0$ such that*

$$\left(1 - \frac{c_1}{\rho_k}\right)\sigma_{k+1} \le \varepsilon_{k+1} + \frac{c_1}{\rho_k}\sigma_k$$

for all $k \in \mathbb{N}$ sufficiently large.

Proof. Observe that $\mathcal{L}_{\rho_k}(x^{k+1}, w^k) = \mathcal{L}(x^{k+1}, \lambda^{k+1})$ for all k. By Assumption 4.2 and the definition of σ_k, we therefore have

$$\sigma_{k+1} \le \varepsilon_{k+1} + \|g(x^{k+1}) - P_K(g(x^{k+1}) + \lambda^{k+1})\|_H. \tag{23}$$

Consider again the sequence $s^{k+1} := P_K(g(x^{k+1}) + \lambda^k/\rho_k)$. Using (22), we see that

$$\|g(x^{k+1}) - P_K(g(x^{k+1}) + \lambda^{k+1})\|_H \le \|g(x^{k+1}) - s^{k+1}\|_H = \frac{\|\lambda^{k+1} - \lambda^k\|_H}{\rho_k}. \tag{24}$$

Inserting this into (23) and using the triangle inequality yields

$$\sigma_{k+1} \le \varepsilon_{k+1} + \frac{1}{\rho_k}\big(\|\lambda^{k+1} - \bar{\lambda}\|_H + \|\lambda^k - \bar{\lambda}\|_H\big).$$

Now, by Assumption 6.2 and since $x^k \to \bar{x}$, there is a $c_1 > 0$ such that $\|\lambda^k - \bar{\lambda}\|_H \le c_1\sigma_k$ for all $k \in \mathbb{N}$ sufficiently large. Hence,

$$\sigma_{k+1} \le \varepsilon_{k+1} + \frac{c_1}{\rho_k}\sigma_{k+1} + \frac{c_1}{\rho_k}\sigma_k,$$

again for $k \in \mathbb{N}$ sufficiently large. Reordering gives the desired result. □

With the above lemma, it is easy to deduce convergence rates for the primal-dual sequence $\{(x^k, \lambda^k)\}$.

Theorem 6.4. *Let Assumptions 4.2, 6.2 hold, and let $\varepsilon_{k+1} = o(\sigma_k)$. Then:*

(a) *For every $q \in (0, 1)$, there is a $\bar{\rho}_q > 0$ such that, if $\rho_k \ge \bar{\rho}_q$ for sufficiently large k, then $(x^k, \lambda^k) \to (\bar{x}, \bar{\lambda})$ Q-linearly with rate q.*

(b) *The sequence of penalty parameters $\{\rho_k\}$ remains bounded.*

Proof. By Lemma 6.3, if ρ_k is large enough so that $1 - c_1/\rho_k > 0$, then

$$\frac{\sigma_{k+1}}{\sigma_k} \le \frac{c_1}{\rho_k - c_1} + o(1). \tag{25}$$

Using (12) and the local Lipschitz-continuity of σ (e.g. equation (14)), it is easy to derive (a). For (b), let us again consider the sequence $s^{k+1} = P_K(g(x^{k+1}) + \lambda^k/\rho_k)$, and define $V_{k+1} := V(x^{k+1}, w^k, \rho_k) = \|\mathcal{L}_{\rho_k}(x^{k+1}, w^k)\|_{X^*} + \|g(x^{k+1}) - s^{k+1}\|_H$. To prove boundedness of $\{\rho_k\}$, we need to show that $V_{k+1} \le \tau V_k$ for sufficiently large k. Using (24) and $\mathcal{L}_{\rho_k}(x^{k+1}, w^k) = \mathcal{L}(x^{k+1}, \lambda^{k+1})$, we obtain

$$V_{k+1} \ge \|\mathcal{L}_{\rho_k}(x^{k+1}, w^k)\|_{X^*} + \|g(x^{k+1}) - P_K(g(x^{k+1}) + \lambda^{k+1})\|_H = \sigma_{k+1}$$

for all $k \in \mathbb{N}$ and, from (24) and Assumption 4.2,

$$\begin{aligned} V_{k+1} = \|\mathcal{L}_{\rho_k}(x^{k+1}, w^k)\|_{X^*} + \frac{\|\lambda^{k+1} - \lambda^k\|_H}{\rho_k} &\le \varepsilon_{k+1} + \frac{\|\lambda^{k+1} - \bar{\lambda}\|_H + \|\lambda^k - \bar{\lambda}\|_H}{\rho_k} \\ &\le \varepsilon_{k+1} + \frac{c}{\rho_k}(\sigma_{k+1} + \sigma_k) \end{aligned}$$

for all $k \in \mathbb{N}$ sufficiently large, where c is the constant from (12) (recall that $x^k \to \bar{x}$). Putting these inequalities together yields

$$\frac{V_{k+1}}{V_k} \le \frac{\varepsilon_{k+1}}{\sigma_k} + \frac{c}{\rho_k} \frac{\sigma_{k+1} + \sigma_k}{\sigma_k} = \frac{\varepsilon_{k+1}}{\sigma_k} + \frac{c}{\rho_k}\left(1 + \frac{\sigma_{k+1}}{\sigma_k}\right).$$

If we now assume that $\rho_k \to \infty$, then it is easy to deduce from (25) and $\varepsilon_{k+1} = o(\sigma_k)$ that $V_{k+1}/V_k \to 0$. Hence, $V_{k+1}/V_k \le \tau$ for all k sufficiently large, which contradicts the assumption that $\rho_k \to \infty$. □

The assumption $\varepsilon_{k+1} = o(\sigma_k)$ in the above theorem says that, roughly speaking, the degree of inexactness should be small enough to not affect the rate of convergence. Note that we are comparing ε_{k+1} to the optimality measure σ_k of the previous iterates (x^k, λ^k). Hence, it is easy to ensure this condition in practice, for instance, by always computing the next iterate x^{k+1} with a precision $\varepsilon_{k+1} \le z_k \sigma_k$ for some fixed zero sequence z_k.

Let us also note that one can easily adapt the proof of Theorem 6.4(a) to conclude that $(x^k, \lambda^k) \to (\bar{x}, \bar{\lambda})$ Q-superlinearly if $\rho_k \to \infty$. However, the resulting assertion would be redundant because part (b) of the theorem actually implies the boundedness of $\{\rho_k\}$. On the other hand, the proof of (b) uses the specific penalty updating scheme (18) with the function V from (16), whereas the proof of (a) does not depend on the penalty updating rule at all. If we replace V by the function

$$\tilde{V}(x, \lambda, \rho) := \left\| g(x) - P_K\left(g(x) + \frac{\lambda}{\rho}\right) \right\|_H$$

(which is just the second term from the definition of V), it is rather easy to see that the assertions of Lemmas 6.1, 6.3 and Theorem 6.4(a) remain true. Additionally, we obtain superlinear convergence if $\rho_k \to \infty$, but we do not get boundedness of $\{\rho_k\}$.

7 Applications

This section describes some applications of our method. Recall that our setting encompasses constrained optimization problems (3). This opens up a broad spectrum of applications, including, of course, standard nonlinear programming (NLP). However, there already is a plethora of literature on this topic, in particular the recent paper [18]. Moreover, the discussion in Section 3 indicates that NLP is actually a very confined special case which does not allow us to demonstrate the full generality of our approach. In particular, NLPs are inherently finite-dimensional and the corresponding set K is polyhedral, which is very restrictive.

As a result, we focus on problems in function space settings where the constraint set is almost never polyhedral. This section contains two examples in this direction: we begin with a simple linear-quadratic optimal control problem and then continue with multiobjective optimal control in a Nash equilibrium framework. Numerical results are provided for both examples.

7.1 An Optimal Control Problem

Let $\Omega \subseteq \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded domain. The example presented in this section is infinite-dimensional and consists of minimizing

$$J(y, u) := \frac{1}{2}\|y - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L^2(\Omega)}^2$$

subject to $y \in H_0^1(\Omega) \cap C(\bar{\Omega})$ and $u \in L^2(\Omega)$ satisfying the partial differential equation (PDE) and pointwise control constraints

$$-\Delta y = u + f \quad \text{and} \quad u_a \leq u \leq u_b.$$

Here, $y_d, u_a, u_b \in L^2(\Omega)$ are problem-specific and $\alpha > 0$ is a regularization parameter. It is well-known that, for every right-hand side $w \in L^2(\Omega)$, the Poisson equation $-\Delta y = w$ admits a uniquely determined weak solution $y = Sw \in H_0^1(\Omega) \cap C(\bar{\Omega})$, and the resulting operator $S : L^2(\Omega) \to H_0^1(\Omega) \cap C(\bar{\Omega})$ is linear and compact [46, Thm. 4.17]. Writing $y_u := S(u + f)$, we can now restate the objective function as

$$\bar{J}(u) := J(y_u, u) = \frac{1}{2}\|y_u - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L^2(\Omega)}^2.$$

This function together with the control constraints $u_a \leq u \leq u_b$ is typically called the *reduced formulation* of the optimal control problem and directly fits into our variational framework by setting $X := H := L^2(\Omega)$, $F(u) := \bar{J}'(u)$, and

$$g(u) := u, \quad K := \{u \in X : u_a \leq u \leq u_b\}.$$

Since $\bar{J}$ is strongly convex and g is just the identity mapping on $X = H$, it is easy to show that the above problem admits a unique primal-dual solution, and that both SOSC and SRC hold. Hence, by Theorem 3.2, the KKT system is upper Lipschitz stable and the control problem admits a local error bound.

	$n=64$			$n=256$			$n=1024$		
k	ρ_k	σ_k	dist_k	ρ_k	σ_k	dist_k	ρ_k	σ_k	dist_k
0	1	5.08e-01	5.43e-01	1	5.02e-01	5.37e-01	1	5.01e-01	5.35e-01
1	1	8.58e-02	1.71e-01	1	8.47e-02	1.69e-01	1	8.44e-02	1.69e-01
2	1	4.29e-02	8.55e-02	1	4.23e-02	8.46e-02	1	4.22e-02	8.44e-02
3	10	2.15e-02	4.26e-02	10	2.12e-02	4.23e-02	10	2.11e-02	4.22e-02
4	10	1.95e-03	3.57e-03	10	1.92e-03	3.83e-03	10	1.92e-03	3.84e-03
5	10	1.77e-04	4.44e-04	10	1.75e-04	3.29e-04	10	1.74e-04	3.48e-04
6	10	1.61e-05	5.08e-04	10	1.59e-05	2.85e-05	10	1.59e-05	3.04e-05
7	10	1.47e-06	5.21e-04	10	1.45e-06	3.18e-05	10	1.44e-06	2.08e-06
8	10	1.33e-07	5.22e-04	10	1.31e-07	3.29e-05	10	1.31e-07	1.96e-06
9	10	1.21e-08	5.22e-04	10	1.20e-08	3.30e-05	10	1.19e-08	2.06e-06
10	10	1.10e-09	5.22e-04	10	1.09e-09	3.30e-05	10	1.08e-09	2.07e-06

Table 1: Numerical results for the optimal control problem.

We now present a numerical example which is constructed in such a way that the optimal solution is known analytically. Let $\Omega := (0,1)^2$ be the unit square and define $\alpha := 1$, $u_a := -0.5$, $u_b := 0.5$. Consider the functions

$$\bar{y}(x) := \sin(\pi x_1)\sin(\pi x_2), \quad \bar{p}(x) := \sin(2\pi x_1)\sin(2\pi x_2),$$

and set $y_d := \bar{y} + \Delta\bar{p}$. Now, using $\bar{u} := P_{[u_a,u_b]}(-\bar{p}/\alpha)$ and $f := -\Delta\bar{y} - \bar{u}$, it is easy to see that $\bar{u}$ is a solution to the problem. Moreover, $\bar{y}$ is the corresponding state, $\bar{p}$ the so-called adjoint state [46], and the Lagrange multiplier is given by $\bar{\lambda} := -\bar{p} - \alpha\bar{u}$.

For the numerical testing, we discretized the problem by means of a uniform grid with $n \in \mathbb{N}$ interior points per row or column (i.e. n^2 points in total) and approximated the Laplace operator by a standard five-point finite difference scheme. The implementation of the algorithm was done in MATLAB® and uses the parameters

$$(u^0, \lambda^0) := (0,0), \quad B := [-10^6, 10^6]^{n^2}, \quad \rho_0 := 1, \quad \gamma := 10, \quad \tau := 0.5,$$

together with the formula $w^k := P_B(\lambda^k)$ for the safeguarded multipliers (see the discussion in Section 4). Moreover, we use the termination criteria $\sigma(x,\lambda) \le 10^{-8}$ and $\|\mathcal{L}_{\rho_k}(x, w^k)\| \le 10^{-10}$ for the outer and inner iterations, respectively, where the norm is the discrete L^2-norm. The subproblems are nonlinear equations which we solve with a standard semismooth Newton method. It should be noted that, while the discrete Laplacian is a sparse matrix, the solution operator S which occurs in the function F is nearly dense. To circumvent this issue, we use a sparse Cholesky factorization of the negative Laplacian to obtain an "implicit" form of S and solve the Newton equations with the MATLAB® conjugate gradient method `pcg`.

Table 1 lists some numerical results for different values of n, where each line contains the penalty parameter ρ_k, the optimality measure σ_k and the distance dist_k of (u^k, λ^k) to $(\bar{u}, \bar{\lambda})$. The results suggest that the algorithm works very well for this problem; in particular, the number of required iterations remains constant as n increases. Moreover, we also observe that the rate of convergence appears to be proportional to $1/\rho_k$, as suggested by the theory. It should be noted, however, that the distances dist_k stop decreasing after a certain point because of the inexactness induced by the discretization; in particular, if we discretize the (known) optimal solution pair $(\bar{u}, \bar{\lambda})$, we do not obtain an *exact* solution of the discretized problem. This phenomenon is also evidenced by the fact that the "limit" value of dist_k decreases as n increases.

We close this section with an important remark on the analytical representation of the feasible set. This observation is crucial and was in fact one of our main motivations to consider constraint sets K which are not necessarily cones.

Remark 7.1. It is important that we define the constraint system with g and K as above. Indeed, the alternative formulation of the box constraints as $\hat{g}(u) \in \hat{K}$ with

$$\hat{g}(u) := (u - u_a, u_b - u), \quad \hat{K} := \{(v, w) \in L^2(\Omega)^2 : v, w \geq 0\},$$

may seem advantageous at first glance (since $\hat{K}$ is a closed convex cone, whereas K is not). However, in this formulation, the strict Robinson condition is not satisfied. In fact, the function $\hat{g}$ does not even satisfy the standard Robinson constraint qualification (RCQ) [9] or the equivalent regularity condition of Zowe and Kurcyusz [50]. We refer the reader to [46] for a formal proof; an alternative way to verify this irregularity is to note that if RCQ holds, then it remains stable under small perturbations of the constraint function [9]. However, even if u_a and u_b are "well separated", it is fairly easy to construct small perturbations (in the sense of L^2) which make the lower and upper bounds coincide on some set of positive measure. If this happens, then the set of Lagrange multipliers corresponding to a local minimum is unbounded, and RCQ is violated.

7.2 Optimal Control in a Nash Equilibrium Framework

We now present a generalization of the optimal control problem from the previous section by considering it in a multi-player framework [10, 14, 35]. The result is a Nash equilibrium problem (NEP) of two players with control variables $u_1, u_2 \in L^2(\Omega)$ and a state variable $y \in H_0^1(\Omega) \cap C(\bar{\Omega})$, where $\Omega \subseteq \mathbb{R}^d$, $d \in \{2, 3\}$, is again a bounded domain. Similarly to before, each player attempts to minimize the objective function

$$J_i(y, u_i) := \frac{1}{2}\|y - y_d^i\|_{L^2(\Omega)}^2 + \frac{\alpha_i}{2}\|u_i\|^2$$

with respect to u_i, subject to the partial differential equation $-\Delta y = u_1 + u_2 + f$ and the pointwise control constraints $a_i \leq u_i \leq b_i$ with $a_i, b_i \in L^2(\Omega)$. The remaining problem parameters satisfy $\alpha_i > 0$ and $y_d^i \in L^2(\Omega)$ for all i. As in Section 7.1, we can use the compact linear solution operator $S : L^2(\Omega) \to H_0^1(\Omega) \cap C(\bar{\Omega})$ and the resulting control-to-state mapping $y_u := S(u_1 + u_2 + f)$ to transform the objective functions to

$$\bar{J}_i(u) := J_i(y_u, u_i) = \frac{1}{2}\|y_u - y_d^i\|_{L^2(\Omega)}^2 + \frac{\alpha_i}{2}\|u_i\|_{L^2(\Omega)}^2,$$

where $u := (u_1, u_2)$. To establish the connection with our variational problem (1), (4), we only need to make some definitions and use the well-known correspondence between NEPs and VIs [15, 35]. Define $X := H := L^2(\Omega)^2$, $F(u) := \big(D_{u_1}\bar{J}_1(u), D_{u_2}\bar{J}_2(u)\big)$, and

$$g(u_1, u_2) := (u_1, u_2), \quad K := \{(u_1, u_2) \in X : a_i \leq u_i \leq b_i\}.$$

k	$n = 64$			$n = 256$			$n = 1024$		
	ρ_k	σ_k	dist_k	ρ_k	σ_k	dist_k	ρ_k	σ_k	dist_k
0	1	5.08e-01	5.43e-01	1	5.02e-01	5.37e-01	1	5.01e-01	5.35e-01
1	1	8.59e-02	1.71e-01	1	8.47e-02	1.69e-01	1	8.44e-02	1.69e-01
2	1	4.30e-02	8.54e-02	1	4.23e-02	8.46e-02	1	4.22e-02	8.44e-02
3	10	2.15e-02	4.24e-02	10	2.12e-02	4.23e-02	10	2.11e-02	4.22e-02
4	10	1.95e-03	3.41e-03	10	1.92e-03	3.81e-03	10	1.92e-03	3.83e-03
5	10	1.78e-04	8.13e-04	10	1.75e-04	3.17e-04	10	1.74e-04	3.47e-04
6	10	1.61e-05	8.95e-04	10	1.59e-05	5.08e-05	10	1.59e-05	2.96e-05
7	10	1.47e-06	9.08e-04	10	1.45e-06	5.63e-05	10	1.44e-06	3.23e-06
8	10	1.33e-07	9.09e-04	10	1.31e-07	5.74e-05	10	1.31e-07	3.50e-06
9	10	1.21e-08	9.09e-04	10	1.20e-08	5.75e-05	10	1.19e-08	3.59e-06
10	10	1.10e-09	9.09e-04	10	1.09e-09	5.75e-05	10	1.08e-09	3.60e-06

Table 2: Numerical results for the optimal control Nash equilibrium problem.

Then it is easy to see that the NEP is equivalent to the VI (1) (and (2), since the feasible set is convex). The existence of a solution of the NEP (and of the VI) can be shown as in [10]; moreover, since g is the identity operator on $X = H$, SRC holds and the problem admits a unique Lagrange multiplier. Finally, an easy calculation shows that

$$F'(u) = \begin{pmatrix} S^*S + \alpha_1 I & S^*S \\ S^*S & S^*S + \alpha_2 I \end{pmatrix},$$

where I is the identity operator on $L^2(\Omega)$. It follows that F is strongly monotone and, since g is linear, the problem automatically satisfies SOSC and therefore admits a local error bound by Theorem 3.2.

We now present some numerical results for the example from [10]. The setting is again constructed in such a way that the optimal solution is known. In fact, the construction is very similar to the one from the previous section: let $\Omega := (0,1)^2$ be the unit square and define $\alpha_i := 1$, $a_i := -0.5$, and $b_i := 0.5$ for all i. Consider the functions

$$\bar{y}(x) := \sin(\pi x_1)\sin(\pi x_2), \qquad \begin{aligned} \bar{p}_1(x) &:= -\sin(2\pi x_1)\sin(2\pi x_2), \\ \bar{p}_2(x) &:= -\sin(3\pi x_1)\sin(3\pi x_2), \end{aligned}$$

as well as $y_d^i := \bar{y} + \Delta p_i$, $\bar{u}_i := P_{[a_i,b_i]}(-\bar{p}_i/\alpha_i)$ for all i, and finally $f := -\Delta\bar{y} - \bar{u}_1 - \bar{u}_2$. Then it is easy to see that $\bar{u}$ is a Nash equilibrium. The corresponding state is given by $\bar{y}$, the variables $\bar{p}_i$ are the adjoint states of the players, and the Lagrange multiplier is given by $\bar{\lambda} := (-\bar{p}_1 - \alpha_1\bar{u}_1, -\bar{p}_2 - \alpha_2\bar{u}_2)$.

The implementation of the augmented Lagrangian method for the above problem is similar to that of the previous section. More precisely, we use the same set of parameters, the same termination criteria, and the same method for the solution of the subproblems. The corresponding numerical results are given in Table 2, where each line contains the values of the penalty parameter ρ_k, the optimality measure σ_k, and the distance dist_k of (u^k, λ^k) to $(\bar{u}, \bar{\lambda})$. We observe good consistency of the results with our established theory; in particular, the rate of convergence is roughly proportional to $1/\rho_k$. We also highlight once again that the distances dist_k do not converge to zero because of the inexactness induced by the discretization.

We close this section by noting that, as explained in Remark 7.1 for the standard (single-objective) optimal control problem, it is very important that we define g and K precisely as we did in order to ensure the fulfillment of the strict Robinson condition.

8 Final Remarks

We have presented a method of augmented Lagrangian type for the solution of variational problems in Banach spaces. In particular, we have shown global and local convergence of the algorithm under suitable assumptions.

The assumptions needed for the local convergence results include, in particular, a local error bound for the distance of a pair (x, λ) to a KKT point $(\bar{x}, \bar{\lambda})$. This property has played a central role in our analysis and is a consequence of the second-order sufficient condition together with a strict version of the Robinson constraint qualification.

The above results lead us to conjecture that error bounds are the natural framework for the local convergence analysis of augmented Lagrangian methods. We therefore hope that the results in this paper will find applications in other areas of optimization. In particular, an interesting idea would be to specialize some of the assumptions and results for problem classes such as optimal control or semidefinite programming. Another aspect which could lead to further developments is the concept of *partial penalization* which arises when additional constraints are present in the problem formulation which are not penalized, see also [1, 6, 8].

References

[1] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt. On augmented Lagrangian methods with general lower-level constraints. *SIAM J. Optim.*, 18(4):1286–1309, 2007.

[2] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt. Augmented Lagrangian methods under the constant positive linear dependence constraint qualification. *Math. Program.*, 111(1-2, Ser. B):5–32, 2008.

[3] C. Baiocchi and A. Capelo. *Variational and Quasivariational Inequalities.* John Wiley & Sons, Inc., New York, 1984.

[4] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer, New York, 2011.

[5] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods.* Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1982.

[6] E. G. Birgin, D. Fernández, and J. M. Martínez. The boundedness of penalty parameters in an augmented Lagrangian method with constrained subproblems. *Optim. Methods Softw.*, 27(6):1001–1024, 2012.

[7] E. G. Birgin, C. A. Floudas, and J. M. Martínez. Global minimization using an augmented Lagrangian method with variable lower-level constraints. *Math. Program.*, 125(1, Ser. A):139–162, 2010.

[8] E. G. Birgin and J. M. Martínez. *Practical Augmented Lagrangian Methods for Constrained Optimization.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2014.

[9] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems.* Springer Series in Operations Research. Springer-Verlag, New York, 2000.

[10] A. Borzì and C. Kanzow. Formulation and numerical solution of Nash equilibrium multiobjective elliptic control problems. *SIAM J. Control Optim.*, 51(1):718–744, 2013.

[11] A. R. Conn, N. I. M. Gould, and P. L. Toint. A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM J. Numer. Anal.*, 28(2):545–572, 1991.

[12] C. Ding, D. Sun, and L. Zhang. Characterization of the robust isolated calmness for a class of conic programming problems. *SIAM J. Optim.*, 27(1):67–90, 2017.

[13] A. L. Dontchev and R. T. Rockafellar. Characterizations of Lipschitzian stability in nonlinear programming. In *Mathematical programming with data perturbations*, volume 195 of *Lecture Notes in Pure and Appl. Math.*, pages 65–82. Dekker, New York, 1998.

[14] A. Dreves and J. Gwinner. Jointly convex generalized Nash equilibria and elliptic multiobjective optimal control. *J. Optim. Theory Appl.*, 168(3):1065–1086, 2016.

[15] F. Facchinei, A. Fischer, and V. Piccialli. On generalized Nash games and variational inequalities. *Oper. Res. Lett.*, 35(2):159–164, 2007.

[16] F. Facchinei and C. Kanzow. Generalized Nash equilibrium problems. *Ann. Oper. Res.*, 175:177–211, 2010.

[17] F. Facchinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems. Vol. I.* Springer-Verlag, New York, 2003.

[18] D. Fernández and M. V. Solodov. Local convergence of exact and inexact augmented Lagrangian methods under the second-order sufficient optimality condition. *SIAM J. Optim.*, 22(2):384–407, 2012.

[19] A. Fischer. Local behavior of an iterative framework for generalized equations with nonisolated solutions. *Math. Program.*, 94(1, Ser. A):91–124, 2002.

[20] A. Fischer, M. Herrich, and K. Schönefeld. Generalized Nash equilibrium problems - recent advances and challenges. *Pesquisa Operacional*, 34:521 – 558, 12 2014.

[21] C. A. Floudas and P. M. Pardalos, editors. *Encyclopedia of Optimization. Vol. I–VI.* Kluwer Academic Publishers, Dordrecht, 2001.

[22] M. Fortin and R. Glowinski. *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, volume 15 of *Studies in Mathematics and its Applications.* North-Holland Publishing Co., Amsterdam, 1983.

[23] R. Glowinski. *Numerical Methods for Nonlinear Variational Problems*. Scientific Computation. Springer-Verlag, Berlin, 2008. Reprint of the 1984 original.

[24] R. Glowinski. *Variational Methods for the Numerical Solution of Nonlinear Elliptic Problems*, volume 86 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2015.

[25] R. Glowinski, J.-L. Lions, and R. Trémolières. *Numerical Analysis of Variational Inequalities*, volume 8 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam-New York, 1981.

[26] M. R. Hestenes. Multiplier and gradient methods. *J. Optimization Theory Appl.*, 4:303–320, 1969.

[27] M. Hintermüller and K. Kunisch. Feasible and noninterior path-following in constrained minimization with low multiplier regularity. *SIAM J. Control Optim.*, 45(4):1198–1221, 2006.

[28] M. Hintermüller, T. Surowiec, and A. Kämmler. Generalized Nash equilibrium problems in Banach spaces: theory, Nikaido-Isoda-based path-following methods, and applications. *SIAM J. Optim.*, 25(3):1826–1856, 2015.

[29] G. Isac. *Complementarity Problems*, volume 1528 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1992.

[30] K. Ito and K. Kunisch. The augmented Lagrangian method for equality and inequality constraints in Hilbert spaces. *Math. Programming*, 46(3, (Ser. A)):341–360, 1990.

[31] K. Ito and K. Kunisch. An augmented Lagrangian technique for variational inequalities. *Appl. Math. Optim.*, 21(3):223–241, 1990.

[32] K. Ito and K. Kunisch. Augmented Lagrangian methods for nonsmooth, convex optimization in Hilbert spaces. *Nonlinear Anal.*, 41(5-6, Ser. A: Theory Methods):591–616, 2000.

[33] K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[34] A. F. Izmailov and M. V. Solodov. Stabilized SQP revisited. *Math. Program.*, 133(1-2, Ser. A):93–120, 2012.

[35] C. Kanzow, V. Karl, D. Steck, and D. Wachsmuth. The multiplier-penalty method for generalized Nash equilibrium problems in Banach spaces. *Technical Report*, Institute of Mathematics, University of Würzburg, July 2017.

[36] C. Kanzow and D. Steck. An example comparing the standard and modified augmented Lagrangian methods. *Technical Report*, Institute of Mathematics, University of Würzburg, February 2017.

[37] C. Kanzow, D. Steck, and D. Wachsmuth. An augmented Lagrangian method for optimization problems in Banach spaces. *Technical Report*, Institute of Mathematics, University of Würzburg, December 2016.

[38] D. Kinderlehrer and G. Stampacchia. *An Introduction to Variational Inequalities and Their Applications*, volume 31 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. Reprint of the 1980 original.

[39] J. Kyparisis. On uniqueness of Kuhn-Tucker multipliers in nonlinear programming. *Math. Programming*, 32(2):242–246, 1985.

[40] J.-J. Moreau. Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires. *C. R. Acad. Sci. Paris*, 255:238–240, 1962.

[41] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, second edition, 2006.

[42] J.-S. Pang and M. Fukushima. Quasi-variational inequalities, generalized Nash equilibria, and multi-leader-follower games. *Comput. Manag. Sci.*, 2(1):21–56, 2005.

[43] M. J. D. Powell. A method for nonlinear constraints in minimization problems. In *Optimization (Sympos., Univ. Keele, Keele, 1968)*, pages 283–298. Academic Press, London, 1969.

[44] R. T. Rockafellar. A dual approach to solving nonlinear programming problems by unconstrained optimization. *Math. Programming*, 5:354–373, 1973.

[45] R. T. Rockafellar. Augmented Lagrange multiplier functions and duality in nonconvex programming. *SIAM J. Control*, 12:268–285, 1974.

[46] F. Tröltzsch. *Optimal Control of Partial Differential Equations*. American Mathematical Society, Providence, RI, 2010.

[47] M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, volume 11 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2011.

[48] G. Wachsmuth. On LICQ and the uniqueness of Lagrange multipliers. *Oper. Res. Lett.*, 41(1):78–80, 2013.

[49] A. P. Wierzbicki and S. Kurcyusz. Projection on a cone, penalty functionals and duality theory for problems with inequality constraints in Hilbert space. *SIAM J. Control Optimization*, 15(1):25–56, 1977.

[50] J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optim.*, 5(1):49–62, 1979.