

Newton-Type Methods: A Broader View

A. F. Izmailov · M. V. Solodov

Received: 27 December 2013 / Accepted: 29 April 2014
© Springer Science+Business Media New York 2014

Abstract We discuss the question of which features and/or properties make a method for solving a given problem belong to the “Newtonian class.” Is it the strategy of linearization (or perhaps, second-order approximation) of the problem data (maybe only part of the problem data)? Or is it fast local convergence of the method under natural assumptions and at a reasonable computational cost of its iteration? We consider both points of view, and also how they relate to each other. In particular, we discuss abstract Newtonian frameworks for generalized equations, and how a number of important algorithms for constrained optimization can be related to them by introducing structured perturbations to the basic Newton iteration. This gives useful tools for local convergence and rate-of-convergence analysis of various algorithms from unified perspectives, often yielding sharper results than provided by other approaches. Specific constrained optimization algorithms, which can be conveniently analyzed within perturbed Newtonian frameworks, include the sequential quadratic programming method and its various modifications (truncated, augmented Lagrangian, composite step, stabilized, and equipped with second-order corrections), the linearly constrained Lagrangian methods, inexact restoration, sequential quadratically constrained quadratic programming, and certain interior feasible directions methods. We recall most of those algorithms as examples to illustrate the underlying viewpoint. We also

Communicated by Aram Arutyunov.

A. F. Izmailov
Moscow State University, MSU, OR Department, VMK Faculty, Uchebniy Korpus 2, Leninskiye Gory,
119991 Moscow, Russia
e-mail: izmaf@ccas.ru

M. V. Solodov (✉)
IMPA – Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico,
Rio de Janeiro, RJ22460-320, Brazil
e-mail: solodov@impa.br

discuss how the main ideas of this approach go beyond clearly Newton-related methods and are applicable, for example, to the augmented Lagrangian algorithm (also known as the method of multipliers), which is in principle not of Newton type since its iterations do not approximate any part of the problem data.

Keywords Newton method · Superlinear convergence · Generalized equation · (Perturbed) Josephy–Newton framework · Constrained optimization · (Perturbed) sequential quadratic programming framework · Augmented Lagrangian

Mathematics Subject Classification 65K05 · 65K10 · 65K15 · 90C30 · 90C33 · 90C53

1 Introduction

The idea of the Newton method is fundamental for solving problems arising in a wide variety of applications. In the setting of nonlinear equations, for which the method was introduced originally, one iteration consists in solving the simpler linear equation, obtained by the first-order approximation of the mapping around the current iterate. If the mapping is smooth enough and the iterative process is initialized from a point close to a solution where the first derivative of the mapping is nondegenerate, then the fast convergence to this solution is guaranteed. Supplied with a suitable technique for globalization of convergence (i.e., for computing an appropriate “starting point”), an efficient algorithm is obtained. Using suitable approximations of the data that defines a given problem, the Newtonian idea can be extended to constrained optimization and to variational problems more general than systems of equations. Moreover, the original differentiability requirements can also be relaxed.

In this article, we survey some classical as well as recent developments in the field of Newton and Newton-type methods. In fact, we take a rather broad view on which methods belong to the Newtonian family. We discuss algorithms which employ first- or second-order approximations to the problem data (and are thus explicitly Newtonian), but also algorithms that can be related to some Newton scheme a posteriori (via perturbations of the Newton iterates) and algorithms that possess fast local convergence under natural assumptions (and are thus related to the Newton method in this sense). Furthermore, the relations between those different points of view are discussed.

The rest of the paper is organized as follows. In Sect. 2, we describe the main problems settings relevant for this article, and the notation employed. In Sect. 3, we discuss some extensions [1] of the main abstract Newtonian iterative frameworks for generalized equations [2–8]), under the three key regularity concepts: strong metric regularity [6], semistability [3], and upper-Lipschitz stability of solutions [4] (some comments on the use of metric regularity are also provided). These results are later applied to more specific algorithms and algorithmic frameworks (we mention that those developments are sometimes not at all immediate, i.e., a fair amount of details may need to be worked out to justify a given application). Section 4 presents the perturbed Josephy–Newton scheme for generalized equations [9] and its stabilized variant [4]. These lead in Sect. 5 to the perturbed sequential quadratic programming framework for optimization [9, 10]

and to stabilized sequential quadratic programming [4, 11–14], respectively. The latter is designed to achieve superlinear convergence for degenerate problems, when no constraint qualifications [15] are assumed, and Lagrange multipliers associated to a given primal solution need not be unique. To illustrate the applications of the perturbed sequential quadratic programming framework, Sect. 5 further discusses truncated and augmented Lagrangian modifications of sequential quadratic programming itself; the linearly constrained Lagrangian methods [16–18], inexact restoration [19–22], and sequential quadratically constrained quadratic programming [23–28]. We note that the framework can also be applied to composite-step versions of sequential quadratic programming [29, 30] or when second-order corrections are used [31], and to certain interior feasible directions methods [32–36]. To keep the length of this exposition reasonable, we shall not discuss these three applications here, referring the readers to [37, 38]. Finally, Sect. 6 puts in evidence that, with appropriate modifications, the Newtonian lines of analysis presented in the preceding parts of the article actually go beyond algorithms that have obvious Newtonian features (i.e., use first- or second-order approximations of the problem data, or at least of some part of the data). The example in this section is the classical augmented Lagrangian algorithm [39–43] (also known as the method of multipliers), which does not use any functions approximations in its iterations.

2 Preliminary Discussion and Problem Statements

Any discussion of Newtonian techniques starts, of course, with the usual *nonlinear* equation

$$\Phi(u) = 0, \tag{1}$$

where $\Phi : \mathbb{R}^v \rightarrow \mathbb{R}^v$ is a given mapping. Let $u^k \in \mathbb{R}^v$ be a given point (e.g., the current iterate of some iterative process for solving (1)). If Φ is differentiable at u^k , then it is natural to approximate (1) near the point u^k by its linearization centred at u^k , i.e., by the *linear* equation

$$\Phi(u^k) + \Phi'(u^k)(u - u^k) = 0. \tag{2}$$

The linearized equation (2) is the iteration subproblem of the classical *Newton method* for solving (1), when Φ is differentiable. Thus, in the basic Newton method, next iterate u^{k+1} is computed by solving (2). The underlying idea is clear—the nonlinear equation (1) is replaced by the (computationally much simpler to solve) linear equation (2). It is expected that (2) can be a good local approximation (at least under some natural assumptions) to the original problem (1), thus giving a rapidly convergent algorithm. Those considerations are formalized in the following classical statement.

Theorem 2.1 *Let $\Phi : \mathbb{R}^v \rightarrow \mathbb{R}^v$ be differentiable in a neighborhood of $\bar{u} \in \mathbb{R}^v$, with its derivative being continuous at \bar{u} . Let \bar{u} be a solution of (1), and assume that $\Phi'(\bar{u})$ is a nonsingular matrix.*

Then, for any starting point $u^0 \in \mathbb{R}^v$ close enough to \bar{u} , there exists a unique sequence $\{u^k\} \subset \mathbb{R}^v$ such that u^{k+1} satisfies (2) for each k , and this sequence converges to \bar{u} superlinearly.

Moreover, if the derivative of Φ is locally Lipschitz-continuous with respect to \bar{u} , the rate of convergence in the assertion of Theorem 2.1 becomes quadratic. In the remainder of this exposition, we shall talk about superlinear convergence only, with the understanding that the quadratic rate can usually be ensured under suitable stronger smoothness assumptions.

One of the main problem settings in this article would be that of the *generalized equation* (GE), which is an inclusion of the form

$$\Phi(u) + N(u) \ni 0, \tag{3}$$

where $\Phi : \mathbb{R}^v \rightarrow \mathbb{R}^v$ is a single-valued base mapping and N is a set-valued field multifunction from \mathbb{R}^v to the subsets of \mathbb{R}^v . If Φ is differentiable, then the *Josephy–Newton method* [44] for the GE (3) is given by the iterative subproblems of the form

$$\Phi(u^k) + \Phi'(u^k)(u - u^k) + N(u) \ni 0, \tag{4}$$

which is an appropriate extension of the Newton iteration (2) to the problem at hand. Under suitable notions of solution regularity of GEs, the scheme described by (4) is locally superlinearly convergent; see Sect. 3.

The GE framework is very broad. For example, the usual system of equations (1) is obtained taking $N(u) = \{0\}$ for all $u \in \mathbb{R}^v$. Another example of a GE is the following *variational problem* (in the terminology of [38]; in [45], this problem is referred to as a “variational condition,” and in some circles, the term “conical relaxed variational inequality” is used):

$$u \in Q, \quad \langle \Phi(u), \xi \rangle \geq 0, \quad \forall \xi \in T_Q(u), \tag{5}$$

where $\Phi : \mathbb{R}^v \rightarrow \mathbb{R}^v$, $Q \subset \mathbb{R}^v$ and $T_Q(u)$ stands for the contingent (Bouligand tangent) cone to the set Q and the point u . The variational problem (5) corresponds to the GE (3) with N being the normal cone multifunction N_Q to the set Q (see the end of this section for the definitions of the contingent and normal cones). If Φ in (5) is the gradient of some differentiable function f , then the variational problem (5) represents the (primal) first-order necessary optimality conditions for u to be a local minimizer of f on the set Q . When the set Q is convex, (5) becomes the *variational inequality* [2,46]

$$u \in Q, \quad \langle \Phi(u), v - u \rangle \geq 0, \quad \forall v \in Q. \tag{6}$$

If the set Q in (6) is a generalized box (some bounds can be infinite), then (6) is the (mixed) complementarity problem.

Another important instance of the GE (3), which is perhaps our main focus in this article, is provided by the primal-dual optimality conditions of constrained optimization problems. To that end, consider the *mathematical programming (or optimization) problem*

$$\text{minimize } f(x) \text{ s.t. } h(x) = 0, \quad g(x) \leq 0, \tag{7}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a smooth objective function, and $h : \mathbb{R}^n \rightarrow \mathbb{R}^l$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are smooth constraint mappings. For the role of the following notions in optimization theory, and their relations to geometric properties coming from separation theorems, we address the readers to [47, 48]. Define the Lagrangian of the problem (7) by $L : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}$,

$$L(x, \lambda, \mu) := f(x) + \langle \lambda, h(x) \rangle + \langle \mu, g(x) \rangle.$$

Stationary points of the problem (7) and the associated Lagrange multipliers are characterized by the *Karush–Kuhn–Tucker (KKT) system*

$$\frac{\partial L}{\partial x}(x, \lambda, \mu) = 0, \quad h(x) = 0, \quad \mu \geq 0, \quad g(x) \leq 0, \quad \langle \mu, g(x) \rangle = 0. \tag{8}$$

The KKT system (8) can be written as the GE (3) with the mapping $\Phi : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ given by

$$\Phi(u) := \left(\frac{\partial L}{\partial x}(x, \lambda, \mu), h(x), -g(x) \right), \tag{9}$$

and with

$$N(\cdot) := N_Q(\cdot), \quad Q := \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}_+^m, \tag{10}$$

where $u = (x, \lambda, \mu)$.

Moreover, it is well known that, for Φ given by (9) and N given by (10), the Josephy–Newton iteration (4) for the GE (3) corresponds to the iteration of the fundamental *sequential quadratic programming* algorithm [49, 50] for the optimization problem (7). Specifically, assuming that the problem data in (7) are twice differentiable, the Josephy–Newton subproblem (4) takes the form

$$\begin{aligned} f'(x^k) + \frac{\partial^2 L}{\partial x^2}(x^k, \lambda^k, \mu^k)(x - x^k) + (h'(x^k))^T \lambda + (g'(x^k))^T \mu &= 0, \\ h(x^k) + h'(x^k)(x - x^k) &= 0, \\ \mu \geq 0, \quad g(x^k) + g'(x^k)(x - x^k) \leq 0, \quad \langle \mu, g(x^k) + g'(x^k)(x - x^k) \rangle &= 0. \end{aligned} \tag{11}$$

It can be easily seen that (11) is the KKT system for the following *quadratic programming problem*, approximating the original optimization problem (7):

$$\begin{aligned} \text{minimize } f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{2} \left\langle \frac{\partial^2 L}{\partial x^2}(x^k, \lambda^k, \mu^k)(x - x^k), x - x^k \right\rangle \\ \text{s.t. } h(x^k) + h'(x^k)(x - x^k) = 0, \quad g(x^k) + g'(x^k)(x - x^k) \leq 0. \end{aligned} \tag{12}$$

The latter is precisely the subproblem of the sequential quadratic programming algorithm, discussed in Sect. 5 below.

The local convergence and rate of convergence properties of the Newton method (2) and its extensions (the Josephy–Newton method (4), sequential quadratic programming (12)) are nice and appealing. However, there exist other algorithms which are not exactly (or sometimes not at all) of the form of the linearization in (2) (partial linearization in (4), the quadratic program in (12)), but which still possess the same or similar local convergence properties as the basic Newton scheme for the corresponding problem. At the same time, those other algorithms may have some advantages compared to the Newton method. For example, computationally cheaper iterations; better global convergence properties; globalization strategies that are more natural/efficient; and weaker assumptions required for the local superlinear convergence (e.g., in the case of Newton method for the usual equation (1), allowing singularity of $\Phi'(\bar{u})$ or even for the solution \bar{u} to be not necessarily isolated).

A natural question arises: which algorithms can be regarded as of Newton type? Is it the form of the iterative subproblems that is crucial (i.e., linearization as in (2), partial linearization as in (4), quadratic programming approximation of the optimization problem as in (12))? Or is it the property of the superlinear convergence guaranteed under natural assumptions and at a reasonable computational cost per iteration (for example, perhaps the optimization problem (7) can be approximated by a subproblem, which is computationally tractable but is not necessarily a quadratic program)? To some extent, this is surely a matter of terminology and convention. Adopting the latter (much wider) viewpoint, many important and successful computational techniques that have nothing or little to do with linearization or quadratic approximations can be seen as Newtonian. Moreover, it is important to emphasize that this approach is often fruitful for gaining a better understanding of the methods in question, or at least a different insight, and in particular for deriving sharper local convergence results than those obtained by other means. This general viewpoint on Newtonian methods is reflected in certain abstract iterative frameworks, such as in [1–7] and in [8, Sect. 6C]. In these frameworks, one explicitly states which properties a method must possess to have superlinear convergence under various sets of assumptions. The other possibility is to go in the opposite direction, that is, to consider specific methods as perturbations of what would be the basic Newton scheme for the given problem. The latter is the central concept adopted in [38]. The goal of this article is to survey both approaches and to relate them to each other.

Our notation is essentially standard; we summarize it next. For two elements u and v in any given space (always finite-dimensional throughout our exposition, and always clear from the context), $\langle u, v \rangle$ stands for the Euclidean inner product between u and v , and $\|\cdot\|$ for the associated norm. By $\min\{u, v\}$, we mean a vector with the minimum taken componentwise (the max operation is analogous). The Euclidean distance from a point u in a given space to a set U in the same space is defined by $\text{dist}(u, U) := \inf_{v \in U} \|u - v\|$. The symbol $B(u, \delta)$ stands for the closed (Euclidean) ball of radius $\delta \geq 0$ centred at u . For an element $u \in \mathbb{R}^v$ and an index set $J \subset \{1, \dots, v\}$, the notation u_J stands for a vector comprised by the components of u indexed by J . Likewise, for a matrix M with v rows, the matrix M_J is comprised by the rows of M with indices in J .

The contingent (Bouligand tangent) cone to the set $Q \subset \mathbb{R}^v$ at a point $u \in Q$ is the set

$$T_Q(u) = \left\{ v \in \mathbb{R}^n \mid \exists \{v^k\} \subset \mathbb{R}^n, \{t_k\} \subset \mathbb{R} \text{ such that } \begin{aligned} &\{v^k\} \rightarrow v, t_k \downarrow 0, u + t_k v^k \in Q, \forall k \end{aligned} \right\}.$$

For a cone $K \subset \mathbb{R}^v$, its polar is $\{v \in \mathbb{R}^v \mid \langle v, u \rangle \leq 0, \forall u \in K\}$. The normal cone $N_Q(u)$ to the set Q at a point $u \in Q$ is the polar of $T_Q(u)$, with the additional convention that $N_Q(u) = \emptyset$ if $u \notin Q$.

For a locally Lipschitz-continuous mapping $\Phi : \mathbb{R}^v \rightarrow \mathbb{R}^v$, its Clarke’s generalized Jacobian at $u \in \mathbb{R}^v$, denoted by $\partial\Phi(u)$, is the convex hull of the B -differential of Φ at u , the latter being the set

$$\{J \in \mathbb{R}^{v \times v} \mid \exists \{u^k\} \text{ in } \mathcal{S}_\Phi \text{ such that } \{u^k\} \rightarrow u, \{\Phi'(u^k)\} \rightarrow J\},$$

where \mathcal{S}_Φ is the set of points at which Φ is differentiable (for Φ locally Lipschitz-continuous, this set is dense, by the Rademacher theorem [51]). Furthermore, for a mapping $F : \mathbb{R}^n \times \mathbb{R}^l \rightarrow \mathbb{R}^m$, the partial generalized Jacobian of F at $(x, y) \in \mathbb{R}^n \times \mathbb{R}^l$ with respect to u is the generalized Jacobian of the mapping $F(\cdot, y)$ at x , which we denote by $\partial_x F(x, y)$.

Let $\mathcal{M}(\bar{x})$ stand for the set of Lagrange multipliers associated to a stationary point \bar{x} of the optimization problem (7), i.e., the set of $(\lambda, \mu) \in \mathbb{R}^l \times \mathbb{R}^m$ satisfying the KKT system (8) for $x = \bar{x}$. Let

$$A(\bar{x}) = \{i = 1, \dots, m \mid g_i(\bar{x}) = 0\}, \quad N(\bar{x}) = \{1, \dots, m\} \setminus A(\bar{x})$$

be the sets of indices of active and inactive constraints at \bar{x} and, for a Lagrange multiplier $(\bar{\lambda}, \bar{\mu}) \in \mathcal{M}(\bar{x})$, let

$$A_+(\bar{x}, \bar{\mu}) = \{i \in A(\bar{x}) \mid \bar{\mu}_i > 0\}, \quad A_0(\bar{x}, \bar{\mu}) = A(\bar{x}) \setminus A_+(\bar{x}, \bar{\mu})$$

be the sets of indices of strongly active and weakly active constraints, respectively.

Recall that the strict Mangasarian–Fromovitz constraint qualification (SMFCQ) at a stationary point \bar{x} of problem (7) consists in saying that the Lagrange multiplier $(\bar{\lambda}, \bar{\mu})$, associated to \bar{x} , exists and is unique. The (in general stronger) linear independence constraint qualification (LICQ) means that the gradients of all the equality constraints, together with the gradients of the active inequality constraints, form a linearly independent set in \mathbb{R}^n .

The second-order sufficient optimality condition (SOSC) holds at \bar{x} for $(\bar{\lambda}, \bar{\mu}) \in \mathcal{M}(\bar{x})$ if

$$\left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda}, \bar{\mu})\xi, \xi \right\rangle > 0, \quad \forall \xi \in C(\bar{x}) \setminus \{0\}, \tag{13}$$

where

$$C(\bar{x}) = \{\xi \in \mathbb{R}^n \mid h'(\bar{x})\xi = 0, g'_{A(\bar{x})}(\bar{x})\xi \leq 0, \langle f'(\bar{x}), \xi \rangle \leq 0\}$$

is the critical cone of problem (7) at \bar{x} . The strong second-order sufficient optimality condition (SSOSC) means that

$$\left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda}, \bar{\mu})\xi, \xi \right\rangle > 0, \quad \forall \xi \in C_+(\bar{x}, \bar{\mu}) \setminus \{0\}, \quad (14)$$

where

$$C_+(\bar{x}, \bar{\mu}) = \{\xi \in \mathbb{R}^n \mid h'(\bar{x})\xi = 0, g'_{A_+(\bar{x}, \bar{\mu})}(\bar{x})\xi = 0\}.$$

3 Abstract Newtonian Iterative Frameworks for Generalized Equations

In this section, we present three abstract iterative frameworks for generalized equations relying on the assumptions of strong metric regularity, semistability, and upper-Lipschitz stability of solutions, respectively (the latter allowing for nonisolated solutions). Some comments regarding metric regularity will also be provided. These frameworks [1] extend the developments in [3,4,6] and in [8, Sect. 6C]. They provide a convenient toolkit for local convergence analysis of various algorithms, under various sets of assumptions. Specific applications will be discussed in the subsequent sections.

To keep the length of this exposition reasonable, when talking about applications of GEs, we shall restrict ourselves to the case of KKT systems (8) of optimization problems (7), with the understanding that there are many other important applications (recall other special cases of GEs mentioned in Sect. 2). We start with a general discussion and survey of the literature, and then move to concrete issues.

Several abstract Newtonian frameworks are well known by now; see [2–7] and [8, Sect. 6C]. These developments can be compared to each other, and to the approach adopted here, by a good number of different criteria. We mention only some key differences. The frameworks in [2,5–7] are designed for solving usual equation (1), while [3,4,8] deal with GEs. The convergence theorems in [4,6–8] require continuity of the equation operator/base mapping; [2,5] employ the assumption of local Lipschitz-continuity; [3] considers continuously differentiable and locally Lipschitzian data (for the latter case, only a posteriori rate of convergence result is given). Each of the schemes mentioned above involves some kind of approximation of the GE base mapping, depending on the current point, which serves as a substitute of the base mapping in the subproblems of the method. In [3,6–8], the approximation is single-valued; in [4,5], it is a multifunction; and in [2], it is a family of mappings.

The local convergence analyses in [2,5–8] rely on some requirements for the quality of the approximation of the GE base mapping in the subproblems, and on appropriate solution regularity. In [3] and [4], solvability of the iteration subproblems is treated separately from regularity of solutions, i.e., solvability is an additional assumption. The solution regularity properties are stated in terms of the approximation in [2,5–7], and in terms of the original base mapping in [3,4,8]. The framework in [5] is much different from all the other schemes. In particular, it allows for solving the subproblems approximately, and this is an essential element of the development therein. However, the framework in [5] does not guarantee that subproblems have (exact) solutions, and therefore the existence of the iterative sequence is not assured (if there are no exact solutions, then there may exist no suitable approximate solutions either).

In [2,4–7], the way the approximation depends on a given point is fixed throughout the process, while in [3,8], it may change from one iteration to another. Here, we shall adopt an even more flexible approach, allowing parametrization of the iterative scheme. This gives more freedom as to what the iteration may depend on, and the resulting abstract frameworks can then be applied to local analyses of algorithms, that may naturally involve parameters. One such example is the augmented Lagrangian method, discussed in Sect. 6.

To that end, for a given set Π of possible parameter values, we consider the class of methods that, given a current iterate $u^k \in \mathbb{R}^v$ and after choosing a parameter value $\pi^k \in \Pi$, generate next iterate u^{k+1} by solving the subproblem of the form

$$\mathcal{A}(\pi^k, u^k, u) + N(u) \ni 0, \tag{15}$$

where, for $\pi \in \Pi$ and $\tilde{u} \in \mathbb{R}^v$, the set-valued mapping $\mathcal{A}(\pi, \tilde{u}, \cdot)$ from \mathbb{R}^v to the subsets of \mathbb{R}^v is some kind of approximation of Φ around \tilde{u} . The required properties of this approximation will be specified below. One obvious example is the Josephy–Newton subproblem (4) above. However, let us emphasize that $\mathcal{A}(\pi^k, u^k, \cdot)$ need not be the standard linearization of Φ at u^k . Moreover, let us further stress that, in the current setting, such a linearization need not even exist, in general, in any reasonable sense (note that differentiability of Φ , or any substitute for it, will not be assumed in this section).

3.1 Strongly Metrically Regular Solutions

We start with the parametric extension of the framework from [8, Sect. 6C], which has its origins in [6].

For each $\pi \in \Pi$ and $\tilde{u} \in \mathbb{R}^v$, define the set

$$U(\pi, \tilde{u}) := \{u \in \mathbb{R}^v \mid \mathcal{A}(\pi, \tilde{u}, u) + N(u) \ni 0\}, \tag{16}$$

so that $U(\pi^k, u^k)$ is the solution set of the iteration subproblem (15). As usual in Newtonian analyses, since even for \tilde{u} arbitrarily close to a solution in question this set may contain points far away, we have to specify which of the solutions of (15) are allowed to be next iterate; solutions “far away” must clearly be discarded from any local analysis (just think of the case when the subproblem means minimizing a nonconvex function: it may have other global and local solutions and stationary points essentially anywhere in the feasible set, in particular far from “the region of interest,” where any local analysis might make sense). Therefore, we have to restrict the distance from the current iterate u^k to next one, i.e., to an element of the set $U(\pi^k, u^k)$ that can be declared to be next iterate u^{k+1} . In the nondegenerate settings (with isolated solutions) of the first two frameworks below, it is sufficient to require that

$$\|u^{k+1} - u^k\| \leq \delta, \tag{17}$$

where $\delta > 0$ is fixed and small enough.

We emphasize that the condition (17) and other localization conditions of this kind cannot be avoided in local convergence analysis for the general problem settings in this article. The reason for this is discussed above; see also Examples 5.1 and 5.2 below, where the issue is illustrated for the SQP algorithm. That said, it must be noted that localization conditions are not directly related to implementations of algorithms. In practice, they are simply ignored, with the understanding that the theoretical assertions apply only if the generated iterates satisfy these conditions. Note, however, that when considering specific algorithms, an important part of the analysis consists precisely in proving that the subproblems in question indeed have solutions satisfying (17). Once we know that such solutions do exist, we can expect that a condition like (17) may naturally hold if the previous iterate u^k is used as a starting point of the process that computes next iterate u^{k+1} . A more direct justification of conditions like (17) would require to consider specific solvers for specific subproblems (which is usually treated as a “black box,” including in this article). This would evidently require a lot of work in each specific case, and perhaps some stronger assumptions, and such developments are certainly beyond the scope of this survey.

Consider the iterative scheme

$$u^{k+1} \in U(\pi^k, u^k) \cap B(u^k, \delta), \quad k = 0, 1, \dots, \tag{18}$$

for some parameters sequence $\{\pi^k\} \subset \Pi$.

Next result assumes that \mathcal{A} is single-valued, and that it approximates Φ in a rather strong sense: the difference $\Phi(\cdot) - \mathcal{A}(\pi, \tilde{u}, \cdot)$ must be locally Lipschitz-continuous at the solution in question, with a sufficiently small Lipschitz constant, uniformly in $\pi \in \Pi$ and in $\tilde{u} \in \mathbb{R}^v$ close to this solution. For a closely related non-parametric result, see [8, Exercise 6C.4].

Theorem 3.1 *Let a mapping $\Phi : \mathbb{R}^v \rightarrow \mathbb{R}^v$ be continuous in a neighborhood of $\bar{u} \in \mathbb{R}^v$, and let N be a set-valued mapping from \mathbb{R}^v to the subsets of \mathbb{R}^v . Let $\bar{u} \in \mathbb{R}^v$ be a solution of the GE (3). Let a set Π and a mapping $\mathcal{A} : \Pi \times \mathbb{R}^v \times \mathbb{R}^v \rightarrow \mathbb{R}^v$ be given.*

Assume that the following properties hold:

- (i) (Strong metric regularity of the solution) *There exists $\ell > 0$ such that, for any $r \in \mathbb{R}^v$ close enough to 0, the perturbed GE*

$$\Phi(u) + N(u) \ni r \tag{19}$$

has near \bar{u} a unique solution $u(r)$, and the mapping $u(\cdot)$ is locally Lipschitz-continuous at 0 with a constant ℓ .

- (ii) (Precision of approximation) *There exists $\bar{\varepsilon} > 0$ such that*
 - (a) $\mathcal{A}(\pi, \tilde{u}, \tilde{u}) = \Phi(\tilde{u})$ for all $\pi \in \Pi$ and all $\tilde{u} \in B(\bar{u}, \bar{\varepsilon})$.
 - (b) *There exists a function $\omega : \Pi \times \mathbb{R}_+^v \times \mathbb{R}^v \times \mathbb{R}^v \rightarrow \mathbb{R}$ such that the inequality $q < 1/2$ holds for*

$$q = \ell \sup \left\{ \omega(\pi, \tilde{u}, u^1, u^2) \mid \pi \in \Pi, \tilde{u}, u^1, u^2 \in B(\bar{u}, \bar{\varepsilon}) \right\},$$

and such that the estimate

$$\begin{aligned} & \|(\Phi(u^1) - \mathcal{A}(\pi, \tilde{u}, u^1)) - (\Phi(u^2) - \mathcal{A}(\pi, \tilde{u}, u^2))\| \\ & \leq \omega(\pi, \tilde{u}, u^1, u^2)\|u^1 - u^2\| \end{aligned} \tag{20}$$

holds for all $\pi \in \Pi$ and all $\tilde{u}, u^1, u^2 \in B(\bar{u}, \bar{\varepsilon})$.

Then, there exist constants $\delta > 0$ and $\varepsilon_0 > 0$ such that, for any starting point $u^0 \in B(\bar{u}, \varepsilon_0)$ and any sequence $\{\pi^k\} \subset \Pi$, there exists a unique sequence $\{u^k\} \subset \mathbb{R}^p$ satisfying (18); this sequence converges to \bar{u} , and for all k it holds that

$$\|u^{k+1} - \bar{u}\| \leq \frac{\ell\omega(\pi^k, u^k, u^k, u^{k+1})}{1 - \ell\omega(\pi^k, u^k, u^k, u^{k+1})}\|u^k - \bar{u}\| \leq \frac{q}{1 - q}\|u^k - \bar{u}\|.$$

In particular, the rate of convergence of $\{u^k\}$ is linear. Moreover, this rate is superlinear if $\omega(\pi^k, u^k, u^k, u^{k+1}) \rightarrow 0$ as $k \rightarrow \infty$.

This specific statement follows from Theorem 3.2 below, and from [52, Theorem 1.4] (the latter essentially says that strong metric regularity is stable subject to single-valued perturbations with a sufficiently small Lipschitz constant).

Theorem 3.1 reveals that the superlinear rate of convergence is achieved if $\Phi(u^{k+1})$ is approximated by $\mathcal{A}(\pi^k, u^k, u^{k+1})$ with progressively better precision as k goes to infinity. Moreover, the tightening of precision can be driven by two different factors: $\omega(\pi^k, u^k, u^k, u^{k+1})$ can be reduced either naturally, as u^k and u^{k+1} approach \bar{u} , or artificially, by an appropriate choice of the parameter values π^k . Examples for both cases will be given below.

If Φ is differentiable near the solution \bar{u} of the GE (3), and its derivative is continuous at \bar{u} , then it follows from the implicit function theorem in [53] that strong metric regularity of \bar{u} is implied by its strong regularity (in fact, under the stated smoothness hypothesis, the two properties are equivalent). Recall that strong regularity consists of saying that the solution \bar{u} of the GE with linearized base mapping

$$\Phi(\bar{u}) + \Phi'(\bar{u})(u - \bar{u}) + N(u) \ni 0$$

is strongly metrically regular. Moreover, this fact was recently extended in [54] to the case when Φ is locally Lipschitz-continuous at the solution \bar{u} but not necessarily smooth: strong metric regularity of \bar{u} is implied by strong regularity of the solution \bar{u} of the GE

$$\Phi(\bar{u}) + J(u - \bar{u}) + N(u) \ni 0$$

for any $J \in \partial\Phi(\bar{u})$. Therefore, in both of these two cases, the assumption (i) of Theorem 3.1 holds automatically.

Remark 3.1 In the case of optimization with sufficiently smooth data, strong (metric) regularity of the solution $\bar{u} = (\bar{x}, \bar{\lambda}, \bar{\mu})$ of the GE associated to the KKT system (8) implies the LICQ, and is implied by the combination of the LICQ with the SSOSC

(14) [53,55]. Moreover, the SSOSC is necessary for strong regularity if \bar{x} is a local solution of the problem (7) [56].

Assumptions (i) and (ii) of Theorem 3.1 will be relaxed in next section, but this will not come for free: we will need to explicitly require solvability of the subproblems, and the iterative sequence will no longer be necessarily unique. Treating solvability of subproblems separately often makes good sense, however, as will be explained in the sequel. The solution regularity assumption will be even further relaxed in our third iterative framework, in particular allowing for nonisolated solutions, but at the price of replacing the localization condition (17) by a stronger one.

3.2 Semistable Solutions

The next result extends the analysis of the Josephy–Newton method for GEs given in [3]. Note that solvability of subproblems is here an assumption, but the regularity condition is weaker than in Sect. 3.1.

Theorem 3.2 *Let a mapping $\Phi : \mathbb{R}^v \rightarrow \mathbb{R}^v$ be continuous in a neighborhood of $\bar{u} \in \mathbb{R}^v$, and let N be a set-valued mapping from \mathbb{R}^v to the subsets of \mathbb{R}^v . Let $\bar{u} \in \mathbb{R}^v$ be a solution of the GE (3). Let \mathcal{A} be a set-valued mapping from $\Pi \times \mathbb{R}^v \times \mathbb{R}^v$ to the subsets of \mathbb{R}^v , where Π is a given set. Assume that the following properties hold:*

- (i) (Semistability of the solution) *There exists $\ell > 0$ such that, for any $r \in \mathbb{R}^v$, any solution $u(r)$ of the perturbed GE (19) close enough to \bar{u} satisfies the estimate*

$$\|u(r) - \bar{u}\| \leq \ell \|r\|.$$

- (ii) (Precision of approximation) *There exist a constant $\bar{\varepsilon} > 0$ and a function $\omega : \Pi \times \mathbb{R}^v \times \mathbb{R}^v \rightarrow \mathbb{R}_+$ such that $q < 1/3$, where*

$$q = \ell \sup \{\omega(\pi, \tilde{u}, u) \mid \pi \in \Pi, \tilde{u}, u \in B(\bar{u}, \bar{\varepsilon}), u \in U(\pi, \tilde{u})\},$$

and the estimate

$$\|w\| \leq \omega(\pi, \tilde{u}, u)(\|u - \tilde{u}\| + \|\tilde{u} - \bar{u}\|) \tag{21}$$

holds for all $w \in \Phi(u) + (-\mathcal{A}(\pi, \tilde{u}, u)) \cap N(u)$, all $\pi \in \Pi$, all $\tilde{u} \in B(\bar{u}, \bar{\varepsilon})$, and all $u \in B(\bar{u}, \bar{\varepsilon})$.

- (iii) (Solvability of subproblems) *For every $\varepsilon > 0$, there exists $\tilde{\varepsilon} > 0$ such that*

$$U(\pi, \tilde{u}) \cap B(\bar{u}, \varepsilon) \neq \emptyset \quad \forall \pi \in \Pi, \forall \tilde{u} \in B(\bar{u}, \tilde{\varepsilon}).$$

Then, there exist some constants $\delta > 0$ and $\varepsilon_0 > 0$ such that, for any starting point $u^0 \in B(\bar{u}, \varepsilon_0)$ and any sequence $\{\pi^k\} \subset \Pi$, the iterative scheme (18) (with an arbitrary choice of u^k satisfying (18)) generates a sequence $\{u^k\} \subset \mathbb{R}^v$; any such sequence converges to \bar{u} , and for all k the following estimate is valid:

$$\|u^{k+1} - \bar{u}\| \leq \frac{2\ell\omega(\pi^k, u^k, u^{k+1})}{1 - \ell\omega(\pi^k, u^k, u^{k+1})} \|u^k - \bar{u}\| \leq \frac{2q}{1 - q} \|u^k - \bar{u}\|. \tag{22}$$

In particular, the rate of convergence of $\{u^k\}$ is linear. Moreover, the rate is superlinear if $\omega(\pi^k, u^k, u^{k+1}) \rightarrow 0$ as $k \rightarrow \infty$.

Strong metric regularity of Sect. 3.1 evidently implies semistability of the solution in question. Observe that unlike in Theorem 3.1, the weaker assumptions of Theorem 3.2 in general do not guarantee uniqueness of the sequence generated by the method.

If one would remove $\|\tilde{u} - \bar{u}\|$ from the right-hand side of the condition (21), then the restriction $q < 1/3$ in Theorem 3.2 can be replaced by $q < 1/2$, and the factor 2 can be removed from the right-hand side of the estimate (22). Then, Theorem 3.1 can be derived employing this modification of Theorem 3.2.

We note that, due to the fact that \mathcal{A} in Theorem 3.2 is allowed to be multivalued, it can be employed to analyze inexact versions of algorithms.

Remark 3.2 In the case of smooth optimization problems, the following characterization of semistability of a solution $\bar{u} = (\bar{x}, \bar{\lambda}, \bar{\mu})$ of the GE, associated to the KKT system (8), has been derived in [3]. Semistability implies the SMFCQ, and is implied by the combination of the SMFCQ with the SOSC (13). Moreover, the SOSC is necessary for semistability if \bar{x} is a local solution of the optimization problem (7). Moreover, semistability implies solvability of the subproblems in the latter case.

We now briefly comment on the line of analysis of iterative methods for GEs that recently emerged from [8, Theorem 6C.1]. This analysis relies on the classical assumption of metric regularity of the solution \bar{u} in question, which consists in saying that there exists $\ell > 0$ such that, for any $r \in \mathbb{R}^v$, the perturbed GE (19) has a solution $u(r)$ satisfying the estimate

$$\text{dist}(u(r), \bar{U}) \leq \ell \|r\|,$$

where \bar{U} is the solution set of the GE (3). Generally, metric regularity is neither weaker nor stronger than semistability. Unlike semistability, metric regularity allows to dispense with the explicit assumption of solvability of subproblems, which is a desirable feature, in principle. However, the price for the convenience to have solvability automatically is actually quite high since, for variational problems, metric regularity is in fact a rather strong requirement. Specifically, as demonstrated in [57], for smooth variational inequalities over polyhedral sets, metric regularity is in fact equivalent to strong regularity. (And smooth variational inequalities over polyhedral sets is perhaps the most important instance of GEs in finite dimensions, subsuming systems of equations with equal number of variables and equations, complementarity problems, KKT systems, etc.) On the other hand, metric regularity allows to consider underdetermined systems of equations, which generically have nonisolated solutions (and in this case, considerations under metric regularity are more related to our Sect. 3.3 below). The facts discussed above lead to the assertion in [8, Theorem 6C.1] reading as follows: for a given solution, there exists an iterative sequence convergent to this solution, without any possibility to characterize how this “good” iterative sequence can be distinguished

from “bad” iterative sequences (to which the assertions do not apply). Recall that in Theorems 3.1 and 3.2, “good” iterative sequences are characterized by the localization condition (17) or (18). Thus, the kind of statement obtained under metric regularity is theoretically interesting and important, but it is in a sense “less practical” and by itself can hardly be regarded as a final local convergence result. To that end, it is complemented in [8, Exercises 6C.3, 6C.4] by the results assuming semistability and strong metric regularity. This gives quite a nice and consistent picture. However, as mentioned above, the statement under strong metric regularity then essentially corresponds to Theorem 3.1 (without parameters), while the statement under semistability, combined with [8, Theorem 6C.1], is not sharper than Theorem 3.2. Generally speaking, metric regularity is perhaps more suitable within analysis of constraint systems than within convergence analysis of algorithms for well-defined variational or optimization problems. For the latter, it seems to make better sense to separate the regularity assumption (e.g., semistability) from the assumption of solvability of subproblems, as is done in Theorem 3.2. The point is that solvability of subproblems can often be established separately for specific algorithms (using their structure), without unnecessarily strengthening the regularity condition in the general framework. This issue will be clear from the applications discussed below.

3.3 Possibly Nonisolated Upper-Lipschitz Stable Solutions

We finally consider an extension of the Newtonian iterative framework for GEs developed in [4]. To tackle the (obvious) difficulties arising when a solution in question is not necessarily isolated, one has to control δ in the localization condition (17), reducing it along the iterations, and at some proper rate. This is the essence of the stabilization mechanism needed in the case of nonisolated solutions to characterize “good” sequences convergent to a solution. Specifically, for an arbitrary but fixed $\sigma > 0$, define the set

$$U^\sigma(\pi, \tilde{u}) := \{u \in U(\pi, \tilde{u}) \mid \|u - \tilde{u}\| \leq \sigma \operatorname{dist}(\tilde{u}, \bar{U})\}, \tag{23}$$

and consider the iterative scheme

$$u^{k+1} \in U^\sigma(\pi^k, u^k), \quad k = 0, 1, \dots \tag{24}$$

Its convergence is given by the following statement.

Theorem 3.3 *Let a mapping $\Phi : \mathbb{R}^v \rightarrow \mathbb{R}^v$ be continuous in a neighborhood of $\bar{u} \in \mathbb{R}^v$, and let N be a set-valued mapping from \mathbb{R}^v to the subsets of \mathbb{R}^v . Let \bar{U} be the solution set of the GE (3), let $\tilde{u} \in \bar{U}$, and assume that the set $\bar{U} \cap B(\tilde{u}, \varepsilon)$ is closed for each $\varepsilon > 0$ small enough. For each k , let \mathcal{A} be a set-valued mapping from $\Pi \times \mathbb{R}^v \times \mathbb{R}^v$ to the subsets of \mathbb{R}^v . Assume that the following properties hold with some fixed $\sigma > 0$:*

- (i) (Upper-Lipschitzian behavior of the solution under canonical perturbations)
There exists $\ell > 0$ such that, for any $r \in \mathbb{R}^v$, any solution $u(r)$ of the per-

turbed GE (19) close enough to \bar{u} satisfies the estimate

$$\text{dist}(u(r), \bar{U}) \leq \ell \|r\|.$$

- (ii) (Precision of approximation) *There exist $\bar{\varepsilon} > 0$ and a function $\omega : \Pi \times \mathbb{R}^v \times \mathbb{R}^v \rightarrow \mathbb{R}_+$ such that $q < 1$, where*

$$q = \ell \sup\{\omega(\pi, \tilde{u}, u) \mid \pi \in \Pi, \tilde{u} \in B(\bar{u}, \bar{\varepsilon}), u \in U^\sigma(\pi, \tilde{u})\},$$

and the estimate

$$\|w\| \leq \omega(\pi, \tilde{u}, u) \text{dist}(\tilde{u}, \bar{U})$$

holds for all $w \in \Phi(u) + (-\mathcal{A}(\pi, \tilde{u}, u)) \cap N(u)$, all $\pi \in \Pi$, all $\tilde{u} \in B(\bar{u}, \bar{\varepsilon})$, and all $u \in \mathbb{R}^v$ satisfying $\|u - \tilde{u}\| \leq \sigma \text{dist}(\tilde{u}, \bar{U})$.

- (iii) (Solvability of subproblems and localization condition) *For any $\pi \in \Pi$ and any $\tilde{u} \in \mathbb{R}^v$ close enough to \bar{u} , the set $U^\sigma(\pi, \tilde{u})$, defined by (16), (23), is nonempty.*

Then, for every $\varepsilon > 0$ there exists $\varepsilon_0 > 0$ such that, for any starting point $u^0 \in B(\bar{u}, \varepsilon_0)$ and any sequence $\{\pi^k\} \subset \Pi$, the iterative scheme (24) (with an arbitrary choice of u^k satisfying (24)) generates a sequence $\{u^k\} \subset B(\bar{u}, \varepsilon)$; any such sequence converges to some $u^* \in \bar{U}$, and for all k , the following estimates are valid:

$$\|u^{k+1} - u^*\| \leq \frac{\sigma \ell \omega(\pi^k, u^k, u^{k+1})}{1 - q} \text{dist}(u^k, \bar{U}) \leq \frac{\sigma q}{1 - q} \text{dist}(u^k, \bar{U}),$$

$$\text{dist}(u^{k+1}, \bar{U}) \leq \ell \omega(\pi^k, u^k, u^{k+1}) \text{dist}(u^k, \bar{U}) \leq q \text{dist}(u^k, \bar{U}).$$

In particular, the rates of convergence of $\{u^k\}$ to u^* and of $\{\text{dist}(u^k, \bar{U})\}$ to zero are superlinear if $\omega(\pi^k, u^k, u^{k+1}) \rightarrow 0$ as $k \rightarrow \infty$.

This iterative framework is the key to the analysis of the stabilized version of the sequential quadratic programming for optimization, and of the stabilized Newton method for variational problems [13, 14]; see Sect. 5.3. Another application is the augmented Lagrangian algorithm [43, 58]; see Sect. 6.

Remark 3.3 In the case of optimization problems (7) with sufficiently smooth data, upper-Lipschitzian behavior of a solution $\bar{u} = (\bar{x}, \bar{\lambda}, \bar{\mu})$ of the GE associated to the KKT system (8) means the following [4, 14, 59]. First, the perturbed GE (19) takes the form of the canonically perturbed KKT system

$$\frac{\partial L}{\partial x}(x, \lambda, \mu) = a, \quad h(x) = b, \quad \mu \geq 0, \quad g(x) \leq c, \quad \langle \mu, g(x) - c \rangle = 0,$$

for $r = (a, b, c) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$. For KKT systems, the upper-Lipschitzian behavior of solutions under canonical perturbations is equivalent to the error bound stating that there exists $\ell > 0$ such that, for all $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, it holds that

$$\|x - \bar{x}\| + \text{dist}((\lambda, \mu), \mathcal{M}(\bar{x})) \leq \ell \left\| \begin{pmatrix} \frac{\partial L}{\partial x}(x, \lambda, \mu) \\ h(x) \\ \min\{\mu, -g(x)\} \end{pmatrix} \right\|. \tag{25}$$

Moreover, this error bound is further equivalent to the so-called noncriticality property of the Lagrange multiplier $(\bar{\lambda}, \bar{\mu})$, which means that there is no triple $(\xi, \eta, \zeta) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$, with $\xi \neq 0$, that satisfies the system

$$\begin{aligned} \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda}, \bar{\mu})\xi + (h'(\bar{x}))^T \eta + (g'(\bar{x}))^T \zeta &= 0, & h'(\bar{x})\xi &= 0, & g'_{A_+(\bar{x}, \bar{\mu})}(\bar{x})\xi &= 0, \\ \zeta_{A_0(\bar{x}, \bar{\mu})} &\geq 0, & g'_{A_0(\bar{x}, \bar{\mu})}(\bar{x})\xi &\leq 0, & \zeta_i(g'_i(\bar{x}), \xi) &= 0, & i \in A_0(\bar{x}, \bar{\mu}), & \zeta_{N(\bar{x})} &= 0. \end{aligned}$$

It is easy to see that noncriticality is implied by the SOSC (13), but is in general a weaker assumption.

4 Perturbed Josephy–Newton Framework

If Φ is differentiable near \bar{u} , with its derivative being continuous at \bar{u} , then Φ can be naturally approximated by its linearization

$$\mathcal{A}(\bar{u}, u) = \Phi(\bar{u}) + \Phi'(\bar{u})(u - \bar{u}),$$

without any parameters. In this case, by the mean-value theorem, assumption (ii) of Theorem 3.1 holds with

$$\omega(\bar{u}, u^1, u^2) = \sup_{t \in [0, 1]} \|\Phi'(tu^1 + (1-t)u^2) - \Phi'(\bar{u})\|,$$

which naturally vanishes as \bar{u}, u^1 , and u^2 tend to \bar{u} . With this particular choice of \mathcal{A} , the iteration subproblem (15) takes the form of that of the Josephy–Newton method (4) for GEs, originally introduced in [44,60]. Theorem 3.1 readily covers the local convergence analysis in [44,60] under the strong regularity assumption. In particular, the statement is as follows.

Theorem 4.1 *Let a mapping $\Phi : \mathbb{R}^v \rightarrow \mathbb{R}^v$ be differentiable in a neighborhood of $\bar{u} \in \mathbb{R}^n$, with its derivative being continuous at \bar{u} , and let N be a set-valued mapping from \mathbb{R}^n to the subsets of \mathbb{R}^n . Let $\bar{u} \in \mathbb{R}^v$ be a strongly regular solution of the GE (3).*

Then, there exists $\delta > 0$ such that, for any starting point $u^0 \in \mathbb{R}^v$ close enough to \bar{u} , there exists a unique sequence $\{u^k\} \subset \mathbb{R}^v$ such that, for each k , the point u^{k+1} is a solution of the GE (4), satisfying the localization condition (17); this sequence converges to \bar{u} , and the rate of convergence is superlinear.

A recent extension of this result to GEs under weaker smoothness assumptions (in particular, with semismooth [61] base mappings) is given in [59]. The iteration

subproblem of this *semismooth Josephy–Newton method* has the form

$$\Phi(u^k) + J_k(u - u^k) + N(u) \ni 0, \quad J_k \in \partial\Phi(u^k), \tag{26}$$

where J_k can be any matrix in the Clarke generalized Jacobian $\partial\Phi(u^k)$. Local super-linear convergence of this method is established in [59] under an appropriate extension of the notion of strong regularity to the nonsmooth case.

We next pass to the important issue of introducing structured perturbations to the Josephy–Newton iteration (4). It is those perturbations that would allow to cast specific algorithms within the given general Newtonian framework. The *perturbed Josephy–Newton method* corresponds to the mapping

$$\mathcal{A}(\tilde{u}, u) = \Phi(\tilde{u}) + \Phi'(\tilde{u})(u - \tilde{u}) + \Omega(\tilde{u}, u - \tilde{u}),$$

where perturbations are characterized by a set-valued mapping Ω .

Theorem 3.2 covers the following more general (than Theorem 4.1) local convergence result obtained in [9, Theorem 2.1] under weaker assumptions.

Theorem 4.2 *Let a mapping $\Phi : \mathbb{R}^v \rightarrow \mathbb{R}^v$ be differentiable in a neighborhood of $\bar{u} \in \mathbb{R}^n$, with its derivative being continuous at \bar{u} , and let N be a set-valued mapping from \mathbb{R}^n to the subsets of \mathbb{R}^n . Let $\tilde{u} \in \mathbb{R}^v$ be a solution of the GE (3). Let Ω be a multifunction from $\mathbb{R}^v \times \mathbb{R}^v$ to the subsets of \mathbb{R}^v . Assume that the following properties hold:*

- (i) (Semistability of the solution) *As in Theorem 3.2.*
- (ii) (Restriction on the perturbations) *The estimate*

$$\omega = o(\|u - \tilde{u}\| + \|\tilde{u} - \bar{u}\|) \tag{27}$$

holds as $\tilde{u} \rightarrow \bar{u}$ and $u \rightarrow \bar{u}$, uniformly for $\omega \in \Omega(\tilde{u}, u - \tilde{u})$, $\tilde{u} \in \mathbb{R}^v$ and $u \in \mathbb{R}^v$ satisfying

$$\Phi(\tilde{u}) + \Phi'(\tilde{u})(u - \tilde{u}) + \omega + N(u) \ni 0.$$

- (iii) (Solvability of subproblems) *For each $\tilde{u} \in \mathbb{R}^v$ close enough to \bar{u} , the GE*

$$\Phi(\tilde{u}) + \Phi'(\tilde{u})(u - \tilde{u}) + \Omega(\tilde{u}, u - \tilde{u}) + N(u) \ni 0 \tag{28}$$

has a solution $u(\tilde{u})$ such that $u(\tilde{u}) \rightarrow \bar{u}$ as $\tilde{u} \rightarrow \bar{u}$.

Then, there exists $\delta > 0$ such that, for any starting point $u^0 \in \mathbb{R}^v$ close enough to \bar{u} , there exists a sequence $\{u^k\} \subset \mathbb{R}^v$ such that, for each k , the point u^{k+1} is a solution of the GE

$$\Phi(u^k) + \Phi'(u^k)(u - u^k) + \Omega(u^k, u - u^k) + N(u) \ni 0, \tag{29}$$

satisfying the localization condition (17); any such sequence converges to \bar{u} , and the rate of convergence is superlinear.

When there are no perturbations (i.e., $\Omega(\cdot) = \{0\}$), Theorem 4.2 recovers the corresponding result for the Josephy–Newton method in [3].

The (perturbed) Josephy–Newton method is a direct extension of the usual Newton method for equations, to the setting of GEs. In particular, linearization is an essential ingredient of the approach, and just as in the classical Newton method, it is expected that its (partially) linearized GE subproblems should be computationally more tractable than the original GE itself, at the same time capturing the local structure under appropriate regularity assumptions.

We note that Theorem 3.3 can also be applied to the (perturbed) Josephy–Newton method, but in the corresponding considerations, the role of perturbations becomes central in the following sense. When the problem is degenerate (by which we mean here that its solutions are possibly nonisolated), one can hardly expect assumption (iii) of Theorem 3.3 (on solvability of subproblems) to hold for the *unperturbed* Josephy–Newton method. Simple examples of usual equations show that, in the degenerate case, the Newton iteration systems (2) need not be solvable no matter how close u^k is to \bar{u} , or when they are solvable, their solutions need not satisfy the estimate $\|u - u^k\| = O(\text{dist}(u^k, \bar{U}))$ as $u^k \rightarrow \bar{u}$. Thus, in the case of degenerate problems, it is the appropriate structural perturbations introduced *intentionally* into the Newton scheme that is the essence of the corresponding algorithm and which “makes it work”; see Sect. 5.3 on stabilized sequential quadratic programming for one illustration. In fact, this gives a good example, showing that perturbations is not necessarily something potentially “harmful” in some sense (like inexact solution of subproblems, which in principle slows the theoretical rate of convergence). Cleverly constructed intentional perturbations of the Newton iteration may actually *help* improve convergence properties of the basic scheme.

Theorem 4.3 *Let a mapping $\Phi : \mathbb{R}^v \rightarrow \mathbb{R}^v$ be differentiable in a neighborhood of $\bar{u} \in \mathbb{R}^n$, with its derivative being continuous at \bar{u} , and let N be a set-valued mapping from \mathbb{R}^n to the subsets of \mathbb{R}^n . Let $\bar{u} \in \mathbb{R}^v$ be a solution of the GE (3). Let Ω be a multi-function from $\mathbb{R}^v \times \mathbb{R}^v$ to the subsets of \mathbb{R}^v . Assume that the following properties hold:*

- (i) (Upper-Lipschitzian behavior of the solutions under canonical perturbations) *As in Theorem 3.3.*
- (ii) (Restriction on the perturbations) *As in Theorem 4.2, but with (27) replaced by*

$$\omega = o(\|u - \bar{u}\|).$$

- (iii) (Solvability of subproblems and localization condition) *For some fixed constant $\sigma > 0$, for each $\bar{u} \in \mathbb{R}^v$ close enough to \bar{u} , the GE (28) has a solution $u(\bar{u})$ such that $\|u(\bar{u}) - \bar{u}\| \leq \sigma \text{dist}(\bar{u}, \bar{U})$.*

Then, for any starting point $u^0 \in \mathbb{R}^v$ close enough to \bar{u} , there exists a sequence $\{u^k\} \subset \mathbb{R}^v$ such that, for each k , the point u^{k+1} is a solution of the GE (29), satisfying the localization condition

$$\|u^{k+1} - u^k\| \leq \sigma \text{dist}(u^k, \bar{U});$$

any such sequence converges to some $u^* \in \bar{U}$, and the rates of convergence of $\{u^k\}$ to u^* and of $\{\text{dist}(u^k, \bar{U})\}$ to zero are superlinear.

We complete this section with the following important observation. Neither Theorem 3.2 nor Theorem 3.3 can be applied to the semismooth Josephy–Newton method (26). Semismoothness cannot guarantee that assumption (ii) of either of those theorems holds when matrices J_k in (26) are chosen as arbitrary elements of $\partial\Phi(u^k)$. That is why the analysis in [59] relies on the extension of strong regularity. In Theorems 4.2 and 4.3, differentiability of the mapping Φ is essential.

5 Applications of Perturbed Josephy–Newton Framework

In this section, we outline some specific constrained optimization algorithms as an illustration of convergence analysis via the perturbed Newtonian frameworks. It is natural to start with the basic sequential quadratic programming method (SQP), which can be cast as the (unperturbed) Josephy–Newton method for the GE associated to the KKT optimality system of the problem. We then introduce certain structured perturbation to the SQP subproblems, and interpret within the resulting framework truncated and augmented Lagrangian modifications of SQP itself, the linearly constrained Lagrangian methods, inexact restoration, and sequential quadratically constrained quadratic programming. We note that the framework can also be used to analyze local convergence of composite-step versions of SQP ([29, 30], see also [62, Sect. 15.4]), SQP with second-order corrections ([31], see also [63, Sect. 17, p. 310], [64, Sect. 18, commentary]), and certain interior feasible directions methods [32–36]. For explanations of the latter three applications, we refer the readers to [37, 38]. We conclude with discussing the stabilized version of SQP, designed for problems with nonisolated solutions.

5.1 Sequential Quadratic Programming

Among the most important specific instances of the Josephy–Newton method for GEs is the SQP algorithm for constrained optimization. As already mentioned in Sect. 2, if the data in problem (7) is twice differentiable and Φ , N are defined according to (9), (10), then the subproblem (4) of the Josephy–Newton method is precisely the KKT optimality system (11) of the QP subproblem (12), which defines the SQP iteration. SQP methods apparently originated from [65], and they are among the most important and widely used tools of computational optimization; see [49, 50] for surveys and historical information. In [3], based on the version of Theorem 4.2 without perturbations, it was established that the SQP method has the local superlinear convergence property assuming the SMFCQ (uniqueness of the Lagrange multiplier) at a stationary point \bar{x} of the problem (7) and the SOSC (13). To this day, this remains the sharpest local convergence result for SQP. Note that the strict complementarity condition $\bar{\mu}_{A(\bar{x})} > 0$ is not assumed, and that in the presence of active inequality constraints SMFCQ is weaker than the LICQ. Strict complementarity and the LICQ are standard (in addition to the SOSC (13)) if other lines of analysis are employed; see, e.g., [63, Theorem 15.2], [64, 66] and [67, Sect. 4.4.3]. In some other previous studies, strict complementarity is dispensed with replacing the SOSC (13) by the SSOSC (14). Note that the LICQ

and the SSOSC imply strong regularity of the solution (recall Remark 3.1), and in this case, convergence and rate of convergence follow from Theorem 4.1, with the additional property that the iteration sequence is *locally* unique.

Theorem 5.1 *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $h : \mathbb{R}^n \rightarrow \mathbb{R}^l$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be twice differentiable in a neighborhood of $\bar{x} \in \mathbb{R}^n$, with their second derivatives being continuous at \bar{x} . Let \bar{x} be a stationary point of problem (7), satisfying the LICQ and the SSOSC (14) for the associated unique Lagrange multiplier $(\bar{\lambda}, \bar{\mu}) \in \mathbb{R}^l \times \mathbb{R}^m$.*

Then, there exists a constant $\delta > 0$ such that, for any starting point $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, there exists a unique sequence $\{(x^k, \lambda^k, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ such that, for each k , the triple $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$ satisfies the system (11), and also satisfies the localization condition

$$\|(x^{k+1} - x^k, \lambda^{k+1} - \lambda^k, \mu^{k+1} - \mu^k)\| \leq \delta; \tag{30}$$

this sequence converges to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, and the rate of convergence is superlinear.

We stress that the localization condition (30) cannot be dropped in Theorem 5.1, even under the additional assumption of strict complementarity. In Example 5.1 below, the original optimization problem (7) has multiple solutions satisfying all the needed assumptions, and convergence to a specific solution cannot be guaranteed without an appropriate localization condition.

Example 5.1 Let $n = 1, l = 0, m = 2, f(x) = x - x^2/2, g(x) = (-x, x - a)$, where $a > 1$ is a given parameter. Problem (7) with this data has two local solutions $\bar{x}^1 = 0$ and $\bar{x}^2 = a$, both satisfying the strongest combination of assumptions: LICQ, strict complementarity, and SOSC (hence, also SSOSC).

Since this is a quadratic programming problem, for any (x^k, μ^k) the corresponding SQP subproblem (12) coincides with the original problem (7); hence, it has the same local solutions \bar{x}^1 and \bar{x}^2 . This shows that no matter how close to a given solution of the original problem one takes (x^k, μ^k) , it is impossible to claim convergence to *this* solution without using localization conditions.

In the example above, the SQP method terminates in one step at *some* solution. Next example demonstrates a more complex and interesting behavior—the SQP iterative sequence may have no solutions at all among its accumulation points, if no localization conditions are employed and arbitrary subproblem solutions are considered.

Example 5.2 Let $n = 1, l = 0, m = 2$, the objective function $f(x) = x$, the constraints mapping $g(x) = (-(x + x^2/2), x + x^2/2 - a)$, where $a > 0$ is a given parameter. Problem (7) with this data has a local solution $\bar{x} = 0$. It can also be made global by adding, e.g., the constraint $x \geq -1$ inactive at \bar{x} . Moreover, in this case, \bar{x} becomes the unique local and global solution. We do not add this constraint here, as this would not change anything for the behavior of the SQP method discussed below. Evidently, \bar{x} satisfies LICQ, the unique associated Lagrange multiplier is $\bar{\mu} = (1, 0)$, and strict complementarity and SOSC (hence, also SSOSC) hold.

For an iterate (x^k, μ^k) , the corresponding SQP subproblem (12) takes the form

$$\text{minimize } x - \frac{1}{2}(\mu_1^k - \mu_2^k)(x - x^k)^2 \quad \text{s.t.} \quad \frac{1}{2}(x^k)^2 \leq (1 + x^k)x \leq a + \frac{1}{2}(x^k)^2.$$

If $a > 1$, and if (x^k, μ^k) is close enough to $(\bar{x}, \bar{\mu})$, then this subproblem has two local solutions: $(x^k)^2/2$ and $a + (x^k)^2/2$, and moreover, the latter will be the only global solution if $a > 2$. Therefore, no matter how close is (x^k, μ^k) to $(\bar{x}, \bar{\mu})$, the SQP method without localization condition can pick $x^{k+1} = a + (x^k)^2/2$ close to a . This demonstrates that one cannot establish convergence to $(\bar{x}, \bar{\mu})$ without localization conditions.

In a Matlab experiment, we observed the following. After jumping from a neighborhood of \bar{x} to a neighborhood of a , the SQP method starts moving back to \bar{x} . Then, once the iterate becomes close enough to \bar{x} , the jump to a neighborhood of a again becomes possible. Such jumps result in a lack of convergence, and moreover, \bar{x} can be not even among the accumulation points of $\{x^k\}$! This behavior is indeed observed, in practice, if the QP solver for subproblems is initialized in a certain very special way. However, if it is initialized in a natural way from the previous iterate, convergence and superlinear rate are observed, of course.

The localization condition can be dropped only under the assumptions guaranteeing that the subproblems have globally unique solutions. This is the case for purely equality-constrained optimization problems, or for problems satisfying some strong convexity assumptions. For general problems with inequality constraints, assumptions that give unique subproblems solutions would be unreasonably strong.

As already commented, convergence of SQP in fact holds under weaker assumptions than those in Theorem 5.1. According to Remark 3.2, the SMFCQ and the SOSC (13) imply assumptions (i) and (iii) of Theorem 4.2. Thus, under the SMFCQ and the SOSC, local convergence result for SQP readily follows from Theorem 4.2. We do not give a formal statement of this result here, since it is a particular case of Theorem 5.2 below (corresponding to having no perturbations therein).

We also note that once the superlinear primal-dual convergence is established or assumed, the only property needed to prove superlinear convergence of the primal part is an error bound condition [68], implied by the SOSC (see Remark 3.3). Recall that in general, superlinear convergence of the primal-dual sequence does not imply any rate for the primal sequence separately [63, Exercise 14.8].

5.2 Perturbed Sequential Quadratic Programming Framework

It can be shown that the SMFCQ and the SOSC guarantee the solvability of SQP-like subproblems, even when they are affected by quite a rich class of perturbations. This fact and Theorem 4.2 were used in [10] to develop a *perturbed SQP* framework and its local convergence theory, conveniently covering a good number of optimization methods, some of which “look very different” from SQP itself. Another important comment is the following. Rather than to directly apply Theorem 4.2 to the methods fitting the perturbed SQP framework, it is much easier to employ Theorem 5.2 below, where the issues of semistability of solutions and solvability of subproblems are already resolved.

To that end, consider the following perturbed version of the SQP iteration subproblem (12):

$$\begin{aligned}
 &\text{minimize } f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{2} \left\langle \frac{\partial^2 L}{\partial x^2}(x^k, \lambda^k, \mu^k)(x - x^k), x - x^k \right\rangle \\
 &\quad + \psi((x^k, \lambda^k, \mu^k), x - x^k) \\
 \text{s.t. } &h(x^k) + h'(x^k)(x - x^k) + \omega_2((x^k, \lambda^k, \mu^k), x - x^k) = 0, \\
 &g(x^k) + g'(x^k)(x - x^k) + \omega_3((x^k, \lambda^k, \mu^k), x - x^k) \leq 0,
 \end{aligned} \tag{31}$$

with some function $\psi : (\mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m) \times \mathbb{R}^n \rightarrow \mathbb{R}$ and some mappings $\omega_2 : (\mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m) \times \mathbb{R}^n \rightarrow \mathbb{R}^l$ and $\omega_3 : (\mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m) \times \mathbb{R}^n \rightarrow \mathbb{R}^m$, which are smooth with respect to the last variable.

Define the function $\Psi : (\mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m) \times (\mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m) \rightarrow \mathbb{R}$,

$$\begin{aligned}
 \Psi((x, \lambda, \mu), (\xi, \eta, \zeta)) := &\psi((x, \lambda, \mu), \xi) + \langle \lambda + \eta, \omega_2((x, \lambda, \mu), \xi) \rangle \\
 &+ \langle \mu + \zeta, \omega_3((x, \lambda, \mu), \xi) \rangle.
 \end{aligned} \tag{32}$$

This function is the Lagrangian of the problem obtained from problem (31) by removing everything except for the perturbation terms. For the elements $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ and $(\xi, \eta, \zeta) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$, set

$$\omega_1((x, \lambda, \mu), (\xi, \eta, \zeta)) := \frac{\partial \Psi}{\partial \xi}((x, \lambda, \mu), (\xi, \eta, \zeta)). \tag{33}$$

Then, the KKT system of the problem (31) takes the following form:

$$\begin{aligned}
 &f'(x^k) + \frac{\partial^2 L}{\partial x^2}(x^k, \lambda^k, \mu^k)(x - x^k) \\
 &+ (h'(x^k))^T \lambda + (g'(x^k))^T \mu + \omega_1((x^k, \lambda^k, \mu^k), (x - x^k, \lambda - \lambda^k, \mu - \mu^k)) = 0, \\
 &h(x^k) + h'(x^k)(x - x^k) + \omega_2((x^k, \lambda^k, \mu^k), x - x^k) = 0, \\
 &\mu \geq 0, \quad g(x^k) + g'(x^k)(x - x^k) + \omega_3((x^k, \lambda^k, \mu^k), x - x^k) \leq 0, \\
 &\langle \mu, g(x^k) + g'(x^k)(x - x^k) + \omega_3((x^k, \lambda^k, \mu^k), x - x^k) \rangle = 0.
 \end{aligned} \tag{34}$$

The terms defined by ω_1, ω_2 , and ω_3 correspond to *structural* perturbations characterizing various specific algorithms for problem (7) within the perturbed SQP framework. These terms contain the information on how a given algorithm differs from SQP, which is considered as the basic Newton method in this context. Choosing (in a meaningful way) the mappings ω_1, ω_2 , and ω_3 , one obtains specific algorithms for solving (7).

It is important to stress that (31) is now not necessarily a QP, although it might be. Since (31) can now be a rather general optimization problem, it makes good sense to consider that only its approximate solution is practical in computational implementations (i.e., some kind of truncation of iterations is necessary within an algorithm applied to solve (31)). To that end, we introduce an additional perturbation that corresponds to the inexactness in solving the subproblems. Note also that even if subproblems are QPs, it can still make good sense to solve them only approximately.

Therefore, we consider that next iterate $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$ satisfies the following inexact version of the (already perturbed) KKT system (34):

$$\begin{aligned} & \left\| f'(x^k) + \frac{\partial^2 L}{\partial x^2}(x^k, \lambda^k, \mu^k)(x - x^k) + (h'(x^k))^T \lambda + (g'(x^k))^T \mu \right. \\ & \left. + \omega_1((x^k, \lambda^k, \mu^k), (x - x^k, \lambda - \lambda^k, \mu - \mu^k)) \right\| \leq \chi_1(x^k, \lambda^k, \mu^k), \\ & \|h(x^k) + h'(x^k)(x - x^k) + \omega_2((x^k, \lambda^k, \mu^k), x - x^k)\| \leq \chi_2(x^k, \lambda^k, \mu^k), \\ & \mu \geq 0, \quad g(x^k) + g'(x^k)(x - x^k) + \omega_3((x^k, \lambda^k, \mu^k), x - x^k) \leq 0, \\ & \langle \mu, g(x^k) + g'(x^k)(x - x^k) + \omega_3((x^k, \lambda^k, \mu^k), x - x^k) \rangle = 0. \end{aligned} \tag{35}$$

Here, $\chi_1 : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ and $\chi_2 : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ are some forcing functions controlling the additional inexactness in solving the subproblems. Note that the conditions in (35) corresponding to the inequality constraints of (31) do not allow for any additional inexactness: these are precisely the corresponding conditions in (34). In the case when inequality constraints are not simple (say, are not bound constraints), this is somewhat of a drawback with respect to solving the subproblems approximately. However, at this time, it is not known how to treat additional inexactness in inequality constraints within this framework, in general.

To make the connection to the perturbed Josephy–Newton scheme, for $u = (x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ and $v = (\xi, \eta, \zeta) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$, set

$$\omega(u, v) = (\omega_1((x, \lambda, \mu), (\xi, \eta, \zeta)), \omega_2((x, \lambda, \mu), \xi), \omega_3((x, \lambda, \mu), \xi)),$$

$$\Theta_1(u) = \{\theta_1 \in \mathbb{R}^n \mid \|\theta_1\| \leq \chi_1(x, \lambda, \mu)\},$$

$$\Theta_2(u) = \{\theta_2 \in \mathbb{R}^l \mid \|\theta_2\| \leq \chi_1(x, \lambda, \mu)\},$$

$$\Theta(u) = \Theta_1(u) \times \Theta_2(u) \times \{0\},$$

where 0 is the zero element in \mathbb{R}^m , and

$$\Omega(u, v) = \omega(u, v) + \Theta(u).$$

The system (35) can then be seen as the iteration GE (29) of the perturbed Josephy–Newton method, where $u^k = (x^k, \lambda^k, \mu^k)$, and Φ and N are defined according to (9) and (10), respectively.

We emphasize that separating the perturbation into the single-valued part $\omega(u, v)$ and the set-valued part $\Theta(u)$ is instructive, because the two parts correspond to perturbations that play conceptually different roles. As explained above, $\omega(u, v)$ represents structural perturbations with respect to the basic SQP iteration, while $\Theta(u)$ stands for additional inexactness allowed when solving the subproblems of the specific method in consideration.

Applying Theorem 4.2, we obtain local convergence properties of perturbed SQP stated in Theorem 5.2 below. The statement is fairly general, but admittedly very

technical. As explained above, it is not a “stand-alone” result, but rather a useful tool for treating various specific algorithms in a unified manner.

Theorem 5.2 *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $h : \mathbb{R}^n \rightarrow \mathbb{R}^l$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be twice differentiable in a neighborhood of $\bar{x} \in \mathbb{R}^n$, with their second derivatives being continuous at \bar{x} . Let \bar{x} be a stationary point of problem (7), satisfying the SMFCQ and the SOS (13) for the associated unique Lagrange multiplier $(\bar{\lambda}, \bar{\mu}) \in \mathbb{R}^l \times \mathbb{R}^m$.*

Let a function $\psi : (\mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m) \times \mathbb{R}^n \rightarrow \mathbb{R}$, and mappings $\omega_2 : (\mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m) \times \mathbb{R}^n \rightarrow \mathbb{R}^l$, $\omega_3 : (\mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m) \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ possess the following properties:

- (i) ψ is continuous at $((\bar{x}, \bar{\lambda}, \bar{\mu}), \xi)$, and $\omega_2(\cdot, \xi)$ and $\omega_3(\cdot, \xi)$ are continuous at $(\bar{x}, \bar{\lambda}, \bar{\mu})$, for every $\xi \in \mathbb{R}^n$ close enough to 0.
- (ii) ψ , ω_2 , and ω_3 are differentiable with respect to ξ in a neighborhood of $((\bar{x}, \bar{\lambda}, \bar{\mu}), 0)$, and twice differentiable with respect to ξ at $((\bar{x}, \bar{\lambda}, \bar{\mu}), 0)$.
- (iii) ω_3 , $\frac{\partial \psi}{\partial \xi}$, $\frac{\partial \omega_2}{\partial \xi}$, and $\frac{\partial \omega_3}{\partial \xi}$ are continuous at $((\bar{x}, \bar{\lambda}, \bar{\mu}), 0)$, and there exists a neighborhood of 0 in \mathbb{R}^n such that $\frac{\partial \omega_2}{\partial \xi}((x, \lambda, \mu), \cdot)$ is continuous on this neighborhood for all $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$.
- (iv) The equalities

$$\begin{aligned} \omega_2((\bar{x}, \bar{\lambda}, \bar{\mu}), 0) &= 0, & \omega_3((\bar{x}, \bar{\lambda}, \bar{\mu}), 0) &= 0, \\ \frac{\partial \psi}{\partial \xi}((\bar{x}, \bar{\lambda}, \bar{\mu}), 0) &= 0, & \frac{\partial \omega_2}{\partial \xi}((\bar{x}, \bar{\lambda}, \bar{\mu}), 0) &= 0, & \frac{\partial \omega_3}{\partial \xi}((\bar{x}, \bar{\lambda}, \bar{\mu}), 0) &= 0 \end{aligned}$$

hold, and for the function Ψ defined by (32), it holds that

$$\left\langle \left(\frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda}, \bar{\mu}) + \frac{\partial^2 \Psi}{\partial \xi^2}((\bar{x}, \bar{\lambda}, \bar{\mu}), (0, 0, 0)) \right) \xi, \xi \right\rangle > 0, \quad \forall \xi \in C(\bar{x}) \setminus \{0\}.$$

Assume further that $\chi_1 : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ and $\chi_2 : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ are any functions satisfying

$$\chi_j(x, \lambda, \mu) = o(\|(x - \bar{x}, \lambda - \bar{\lambda}, \mu - \bar{\mu})\|), \quad j = 1, 2;$$

and that the estimates

$$\omega_j((x, \lambda, \mu), \xi) = o(\|\xi\| + \|(x - \bar{x}, \lambda - \bar{\lambda}, \mu - \bar{\mu})\|), \quad j = 2, 3,$$

$$\frac{\partial \Psi}{\partial \xi}((x, \lambda, \mu), (\xi, \eta, \zeta)) = o(\|(\xi, \eta, \zeta)\| + \|(x - \bar{x}, \lambda - \bar{\lambda}, \mu - \bar{\mu})\|)$$

hold as the point $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ tends to $(\bar{x}, \bar{\lambda}, \bar{\mu})$ and as the element $(\xi, \eta, \zeta) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ tends to zero, for (x, λ, μ) and (ξ, η, ζ) satisfying

$$\begin{aligned} & \left\| f'(x) + \frac{\partial^2 L}{\partial x^2}(x, \lambda, \mu)\xi + (h'(x))^T(\lambda + \eta) + (g'(x))^T(\mu + \zeta) \right. \\ & \quad \left. + \frac{\partial \Psi}{\partial \xi}((x, \lambda, \mu), (\xi, \eta, \zeta)) \right\| \leq \chi_1(x, \lambda, \mu), \\ & \|h(x) + h'(x)\xi + \omega_2((x, \lambda, \mu), \xi)\| \leq \chi_2(x, \lambda, \mu), \\ & \mu + \zeta \geq 0, \quad g(x) + g'(x)\xi + \omega_3((x, \lambda, \mu), \xi) \leq 0, \\ & (\mu + \zeta, g(x) + g'(x)\xi + \omega_3((x, \lambda, \mu), \xi)) = 0. \end{aligned}$$

Then, there exists a constant $\delta > 0$ such that, for any starting point $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, there exists a sequence $\{(x^k, \lambda^k, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ such that, for each k , the triple $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$ satisfies the system (35) with ω_1 defined in (33) and Ψ defined in (32), and also satisfies the localization condition (30); any such sequence converges to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, and the rate of convergence is superlinear.

In what follows, we discuss some applications of Theorem 5.2. As mentioned in the beginning of this section, these are just some of the examples; the list below is not exhaustive. We emphasize, however, that application of this theorem and other general local convergence results presented above to specific algorithms is not always straightforward: sometimes, verification of the assumptions of these results may require quite a lot of effort.

5.2.1 Augmented Lagrangian and Truncated Modifications of Sequential Quadratic Programming

To illustrate the applications of perturbed SQP framework, we shall start with two useful modifications of the basic SQP itself: using the augmented Lagrangian (instead of the usual Lagrangian) to define the matrix in the objective function of the SQP subproblem, and truncating the solution of subproblems (i.e., solving subproblems approximately). Even though these two issues are, in principle, independent, it is easy for us to deal with them simultaneously.

The development of truncated SQP methods for equality-constrained problems is not too difficult, at least conceptually, because the optimality conditions of the subproblems contain equations only, and thus the ideas of truncated Newton methods for equations (e.g., [69]) naturally extend. The case of inequality constraints, on the other hand, leads to some principal difficulties; see the discussion in [70,71]. If the iteration QP (12) is solved by some finite active-set method, then the truncation is hardly possible. Indeed, none of the iterates produced by an active-set QP method, except for the very last one (at termination), can be expected to approximate the QP solution in any reasonable sense (the algorithm “jumps” to a solution once the working active set is correct, rather than approaches a solution asymptotically). An alternative is to solve QPs approximately by some interior-point method, as suggested, e.g., in [10,72]. This approach is justifiable for the specific issue of truncation, although

there are still some practical challenges to deal with (e.g., warm-starting interior-point methods is not straightforward). Speaking about the theoretical side, for local analysis specifically within the perturbed SQP framework, the difficulty is the necessity to avoid any truncation for the part of optimality conditions of the QP that involve the inequality constraints; see (35) and the discussion following it. As interior-point methods do not maintain complementarity (perturbing complementarity is, in fact, the essence of this class of algorithms), conditions like the last two lines in (35) do not hold along the iterations of solving the QP subproblem.

An approach, that resolves the outlined difficulty, can be constructed for the special but important case of bound constraints. To that end, consider the problem (7) with equality constraints and simple (non-negativity) bounds only, the latter represented by the inequality constraint mapping $g(x) = -x, x \in \mathbb{R}^n$. Therefore, the problem is

$$\text{minimize } f(x) \text{ s.t. } h(x) = 0, x \geq 0. \tag{36}$$

Let $L : \mathbb{R}^n \times \mathbb{R}^l \rightarrow \mathbb{R}$ be the (partial) Lagrangian of problem (36), including only the equality constraints:

$$L(x, \lambda) = f(x) + \langle \lambda, h(x) \rangle.$$

We consider the truncated version of SQP allowing at the same time for the augmented Lagrangian regularization of the matrix in the objective function of the SQP subproblem. Specifically, we set

$$H_k = \frac{\partial^2 L}{\partial x^2}(x^k, \tilde{\lambda}^k) + c(h'(x^k))^T h'(x^k), \tag{37}$$

where $c \geq 0$ is the penalty parameter, and $\tilde{\lambda}^k$ is generated by the rule

$$\tilde{\lambda}^0 = \lambda^0, \quad \tilde{\lambda}^k = \lambda^k - ch(x^{k-1}), \quad k = 1, 2, \dots \tag{38}$$

As is well known, this matrix H_k is closely related to the Hessian of the augmented Lagrangian function (see, e.g., the definition in Sect. 6 below). The motivation for this modification is that H_k , given by (37), has a much better chance of being positive definite than the Hessian of the usual Lagrangian, which is important for efficient/reliable solution of subproblems, as well as for certain globalization SQP strategies based on nonsmooth penalty functions; e.g., [63, Chap. 17]. Note, however, that the value $c = 0$ corresponding to the usual Lagrangian (and thus giving the usual SQP) is also allowed in our framework.

The *truncated SQP method* for problem (36) generates next primal-dual iterate $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$ by computing a triple satisfying the system

$$\begin{aligned} \|f'(x^k) + H_k(x - x^k) + (h'(x^k))^T \lambda - \mu\| &\leq \varphi(\rho(x^k, \tilde{\lambda}^k, \mu^k)), \\ \|h(x^k) + h'(x^k)(x - x^k)\| &\leq \varphi(\rho(x^k, \tilde{\lambda}^k, \mu^k)), \\ \mu &\geq 0, \quad x \geq 0, \quad \langle \mu, x \rangle = 0, \end{aligned} \tag{39}$$

where H_k and $\tilde{\lambda}^k$ are given by (37) and (38) with some $c \geq 0$; $\varphi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a forcing function, and $\rho : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n \rightarrow \mathbb{R}_+$ is the natural residual for optimality conditions of the problem (36), i.e.,

$$\rho(x, \lambda, \mu) = \left\| \left(\frac{\partial L}{\partial x}(x, \lambda) - \mu, h(x), \min\{\mu, x\} \right) \right\|.$$

Note that the set of relations (39) is an inexact KKT system of the QP

$$\begin{aligned} &\text{minimize } f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{2} \langle H_k(x - x^k), x - x^k \rangle \\ \text{s.t. } & \quad h(x^k) + h'(x^k)(x - x^k) = 0, \quad x \geq 0. \end{aligned} \tag{40}$$

Some comments about the use of interior-point methods within truncated SQP are in order. Primal-dual interior-point methods applied to the QP (40) generate a sequence of points $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ satisfying $x > 0$ and $\mu > 0$, and therefore the last line of relations in (39) can never hold except in the limit. The following simple purification procedure resolves this problem: for (x, λ, μ) produced by an iteration of the given interior-point method, define auxiliary points $\hat{x} \in \mathbb{R}^n$ and $\hat{\mu} \in \mathbb{R}^n$ (automatically satisfying the last line in (39)) by

$$\hat{x}_j := \begin{cases} x_j, & \text{if } x_j \geq \mu_j, \\ 0, & \text{if } x_j < \mu_j, \end{cases} \quad \hat{\mu}_j := \begin{cases} 0, & \text{if } x_j \geq \mu_j, \\ \mu_j, & \text{if } x_j < \mu_j, \end{cases} \quad j = 1, \dots, n,$$

and verify (39) for $(\hat{x}, \lambda, \hat{\mu})$. If (39) is satisfied, then accept $(\hat{x}, \lambda, \hat{\mu})$ as the new iterate of the truncated SQP method. Otherwise proceed with next iteration of the interior-point method for the QP (40) to obtain next (x, λ, μ) , perform purification again, verify (39) for the resulting purified point, etc. Under standard assumptions, the interior-point method drives (x, λ, μ) to the solution of the QP (40), and it can be seen that the purified $(\hat{x}, \lambda, \hat{\mu})$ then converges to the same limit. Therefore, next iterate satisfying (39) is obtained after a finite number of interior-point iterations if $\varphi(\rho(x^k, \tilde{\lambda}^k, \mu^k)) > 0$; see [10] for details.

We next explain how the described truncated SQP with augmented Lagrangian modification fits perturbed SQP framework. Observe that (39) is precisely the perturbed SQP iteration subproblem (35) with λ^k replaced by $\tilde{\lambda}^k$, with the variable λ shifted by $-ch(x^k)$, and with

$$\omega_1((x, \lambda, \mu), (\xi, \eta, \zeta)) = \omega_1(x, \xi) = c(h'(x))^T(h(x) + h'(x)\xi),$$

$$\omega_2((x, \lambda, \mu), \xi) = 0, \quad \omega_3((x, \lambda, \mu), \xi) = 0$$

for $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ and $\xi \in \mathbb{R}^n$, and with appropriate χ_1 and χ_2 . Moreover, (39) is the “truncated” KKT system for the perturbed SQP subproblem (31), where one takes

$$\psi((x, \lambda, \mu), \xi) = \psi(x, \xi) = \frac{c}{2} \|h(x) + h'(x)\xi\|^2.$$

With these definitions of ψ , ω_1 , ω_2 , and ω_3 , it holds that Ψ coincides with ψ (see (32)), and the equality (33) is valid.

Applying Theorem 5.2 with the appropriate $\chi_1 = \chi_2$ (and with λ^k substituted by $\tilde{\lambda}^k$!), we derive local convergence properties of the truncated SQP method with (or without, if $c = 0$) the augmented Lagrangian modification of the QP matrix.

Theorem 5.3 *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $h : \mathbb{R}^n \rightarrow \mathbb{R}^l$ be twice differentiable in a neighborhood of $\bar{x} \in \mathbb{R}^n$, with their second derivatives continuous at \bar{x} . Let \bar{x} be a stationary point of problem (36), satisfying the SMFCQ and the SOSC for the associated unique Lagrange multiplier $(\bar{\lambda}, \bar{\mu}) \in \mathbb{R}^l \times \mathbb{R}^n$.*

Then, for any fixed $c \geq 0$, there exists $\delta > 0$ such that, for any function $\varphi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $\varphi(t) = o(t)$ as $t \rightarrow 0$, and for any starting point $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, there exists a sequence $\{(x^k, \lambda^k, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ such that, for each k , the triple $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$ satisfies (39) with H_k and $\tilde{\lambda}^k$ defined by (37) and (38), and also satisfies the localization condition (30); any such sequence converges to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, the sequence $\{\tilde{\lambda}^k\}$ converges to $\bar{\lambda}$, and the rate of convergence of $\{(x^k, \tilde{\lambda}^k, \mu^k)\}$ is superlinear.

5.2.2 Linearly Constrained Lagrangian Methods

As already mentioned above, it is important to stress that algorithms that can be analyzed via perturbed SQP framework are not restricted to those whose iterations consist in solving QP subproblems. A good example is provided by the *linearly constrained (augmented) Lagrangian methods* (LCL); see, e.g., [16–18]. These methods are traditionally stated for problems with equality and bound constraints. Therefore, as in the previous section, we consider the problem in the format of (36).

An iteration subproblem of LCL for (36), defining next primal iterate, consists in minimizing the (augmented) Lagrangian subject to the linearized equality constraints and bounds:

$$\text{minimize } L(x, \lambda^k) + \frac{c_k}{2} \|h(x)\|^2 \quad \text{s.t. } h(x^k) + h'(x^k)(x - x^k) = 0, \quad x \geq 0. \quad (41)$$

Let η^k be a Lagrange multiplier associated to a stationary point x^{k+1} of this subproblem. Then, next dual iterate is given by $\lambda^{k+1} = \lambda^k + \eta^k$.

In the original LCL method proposed in [16], the subproblems involve the usual Lagrangian rather than the augmented Lagrangian, i.e., $c_k = 0$ for all k . However, in practice, it is often important to employ $c_k > 0$; see [73].

The KKT system of the LCL subproblem (41) has the form

$$\begin{aligned} \frac{\partial L}{\partial x}(x, \lambda^k) + c_k (h'(x))^T h(x) + (h'(x^k))^T \eta - \mu &= 0, \\ h(x^k) + h'(x^k)(x - x^k) &= 0, \\ \mu \geq 0, \quad x \geq 0, \quad \langle \mu, x \rangle &= 0, \end{aligned} \quad (42)$$

with the dual variables $\eta \in \mathbb{R}^l$ and $\mu \in \mathbb{R}^n$. It turns out that the subproblem (41) does not fit the assumptions of Theorem 5.2 directly. However, as suggested in [10], there is

a simple theoretically equivalent transformation that makes Theorem 5.2 applicable. Specifically, consider the following transformation of the subproblem (41), which is equivalent to it:

$$\begin{aligned} &\text{minimize } L(x, \lambda^k) - \langle \lambda^k, h'(x^k)(x - x^k) \rangle + \frac{c_k}{2} \|h(x)\|^2 \\ &\text{s.t. } \quad h(x^k) + h'(x^k)(x - x^k) = 0, \quad x \geq 0. \end{aligned} \tag{43}$$

(We emphasize that the LCL method still solves (41), of course, but its solutions and stationary points are in the obvious correspondence with solutions and stationary points of (43), and these would be used for the purposes of convergence analysis.)

The KKT system of the modified LCL subproblem (43) is given by

$$\begin{aligned} \frac{\partial L}{\partial x}(x, \lambda^k) + c_k (h'(x))^T h(x) + (h'(x^k))^T (\lambda - \lambda^k) - \mu &= 0, \\ h(x^k) + h'(x^k)(x - x^k) &= 0, \\ \mu \geq 0, \quad x \geq 0, \quad \langle \mu, x \rangle &= 0, \end{aligned} \tag{44}$$

with the dual variables $\lambda \in \mathbb{R}^l$ and $\mu \in \mathbb{R}^n$. Comparing (42) and (44), we observe that stationary points x^{k+1} of problems (41) and (43) coincide, and the associated multipliers are of the form (η^k, μ^{k+1}) and $(\lambda^{k+1}, \mu^{k+1})$, with $\lambda^{k+1} = \lambda^k + \eta^k$. Thus, for the purposes of convergence analysis, we can deal with the modified subproblems (43), and it turns out that this does allow to apply Theorem 5.2.

For asymptotic analysis, we may consider that c_k is fixed at some value $c \geq 0$ for all k sufficiently large; this happens for typical penalty parameters update rules under natural assumptions (see, e.g., the discussion in [18]).

Note that the constraints of the LCL subproblem (43) are exactly the same as in the basic SQP subproblem (12), once the latter is specialized for the current problem setting of (36). Thus, the structural perturbation that defines the LCL method within the perturbed SQP framework is induced by the (different, non-quadratic) objective function in (43). It can be seen that the LCL subproblem (43) is a particular case of the perturbed SQP subproblem (31), where for $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ and $\xi \in \mathbb{R}^n$ one takes

$$\begin{aligned} \psi((x, \lambda, \mu), \xi) &= \psi((x, \lambda), \xi) \\ &= L(x + \xi, \lambda) - \langle \lambda, h'(x)\xi \rangle + \frac{c}{2} \|h(x + \xi)\|^2 \\ &\quad - f(x) - \langle f'(x), \xi \rangle - \frac{1}{2} \left\langle \frac{\partial^2 L}{\partial x^2}(x, \lambda)\xi, \xi \right\rangle \\ &= L(x + \xi, \lambda) - \left\langle \frac{\partial L}{\partial x}(x, \lambda), \xi \right\rangle - \frac{1}{2} \left\langle \frac{\partial^2 L}{\partial x^2}(x, \lambda)\xi, \xi \right\rangle \\ &\quad + \frac{c}{2} \|h(x + \xi)\|^2 - f(x), \end{aligned} \tag{45}$$

$$\omega_2((x, \lambda, \mu), \xi) = 0, \quad \omega_3((x, \lambda, \mu), \xi) = 0. \tag{46}$$

Employing the mean-value theorem, one can directly verify that under the appropriate smoothness assumptions on f and h , the function ψ and the mappings ω_2 and ω_3 , defined by (45), (46), possess all the properties required in Theorem 5.2 with $\chi_1(\cdot) \equiv 0$ and $\chi_2(\cdot) \equiv 0$. Hence, we have the following local convergence properties for the LCL methods.

Theorem 5.4 *Under the assumptions of Theorem 5.3, for any fixed constant $c \geq 0$, there exists a constant $\delta > 0$ such that, for any starting point $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, there exists a sequence $\{(x^k, \lambda^k, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ such that, for each k , the point x^{k+1} is a stationary point of problem (41) and $(\lambda^{k+1} - \lambda^k, \mu^{k+1})$ is an associated Lagrange multiplier, satisfying the localization condition (30); any such sequence converges to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, and the rate of convergence is superlinear.*

We note that the previous literature required, in addition to the assumptions above, the stronger LICQ and the strict complementarity condition [16, 17]. Moreover, as in the case of SQP, the fact that the primal part of the sequence also converges superlinearly, again follows from [68].

Taking into account that (41) is now not a QP, but a more general problem, it makes good practical sense to allow for inexact solution of the LCL subproblems. Consider the *truncated LCL method*, replacing the system (42) by a version where the parts corresponding to general nonlinearities are relaxed:

$$\begin{aligned} \left\| \frac{\partial L}{\partial x}(x, \lambda^k) + c_k(h'(x))^T h(x) + (h'(x^k))^T \eta - \mu \right\| &\leq \varphi(\rho(x^k, \lambda^k, \mu^k)), \\ \|h(x^k) + h'(x^k)(x - x^k)\| &\leq \varphi(\rho(x^k, \lambda^k, \mu^k)), \\ \mu \geq 0, \quad x \geq 0, \quad \langle \mu, x \rangle &= 0, \end{aligned} \tag{47}$$

with some forcing function $\varphi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$. Local convergence properties of the truncated LCL method also follow from Theorem 5.2 with the appropriate $\chi_1 = \chi_2$.

Theorem 5.5 *Under assumptions of Theorem 5.3, for any fixed $c \geq 0$, there exists $\delta > 0$ such that, for any function $\varphi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $\varphi(t) = o(t)$ as $t \rightarrow 0$, and for any starting point $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, there exists a sequence $\{(x^k, \lambda^k, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ such that, for each k , the triple $(x^{k+1}, \lambda^{k+1} - \lambda^k, \mu^{k+1})$ satisfies (47) and the localization condition (30); any such sequence converges to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, and the rate of convergence is superlinear.*

5.2.3 Inexact Restoration Methods

Inexact restoration methods are also traditionally stated for problems with equality and bound constraints. Therefore, our problem setting in this section remains that of (36). We shall consider inexact restoration methods along the lines of the local framework in [20]. For some other references on methods of this class see [19, 21, 22, 74], and [75] for a survey. The approach to cast inexact restoration within perturbed SQP was proposed in [37].

It is instructive to start the discussion with the “exact” restoration method, which is not a practical algorithm, but rather a motivation for inexact restoration. In particular, in the terminology of this article, it is the exact restoration that defines structural perturbations with respect to SQP, while the inexactness feature is naturally handled as truncation of the perturbed SQP subproblems.

Each iteration of the conceptual exact restoration scheme has two phases. First, on the feasibility phase, π^k is computed as a projection of x^k onto the feasible set of the problem (36), i.e., π^k is a global solution of the subproblem

$$\text{minimize } \|\pi - x^k\| \text{ s.t. } h(\pi) = 0, \quad \pi \geq 0. \tag{48}$$

Then, on the optimality phase, next primal iterate x^{k+1} and the pair (η^k, μ^{k+1}) are computed as a stationary point and an associated Lagrange multiplier of the subproblem

$$\text{minimize } L(x, \lambda^k) \text{ s.t. } h'(\pi^k)(x - \pi^k) = 0, \quad x \geq 0. \tag{49}$$

Next, dual iterate corresponding to equality constraints is then given by $\lambda^{k+1} = \lambda^k + \eta^k$.

Similar to the LCL method in Sect. 5.2.2, (in)exact restoration does not directly fit the perturbed SQP framework in terms of the assumptions imposed on it in Theorem 5.2, but there is an equivalent transformation that does the job: replace the optimality phase subproblem (49) by the following:

$$\text{minimize } L(x, \lambda^k) - \langle \lambda^k, h'(\pi^k)(x - x^k) \rangle \text{ s.t. } h'(\pi^k)(x - \pi^k) = 0, \quad x \geq 0. \tag{50}$$

It is easily seen that stationary points x^{k+1} of the problems (49) and (50) coincide, and the associated multipliers are (η^k, μ^{k+1}) in (49) and $(\lambda^{k+1}, \mu^{k+1})$ in (50), where $\lambda^{k+1} = \lambda^k + \eta^k$. Thus, for the purposes of convergence analysis, one can deal with the modified subproblems (50).

For a given $x \in \mathbb{R}^n$, let $\bar{\pi}(x)$ be a projection of x onto the feasible set of the problem (36), computed as at the feasibility phase of the method for $x = x^k$. Note that to be able to formally apply Theorem 5.2, $\bar{\pi}(\cdot)$ has to be a fixed single-valued function. As the feasible set here is not convex, the projection onto it need not be unique. Nevertheless, an algorithm used to solve (48) follows its internal rules and iterations, and computes one specific projection. It is clearly reasonable to assume that, if at some future iteration a projection of the same point needs to be computed (which is already a highly unlikely event), then the (same) algorithm used for this task would return the same result (even if the projection is not unique). Thus, the assumption that $\bar{\pi}(\cdot)$ is a single-valued function is realistic for computational practice. In Theorem 5.6 below, this assumption is stated more formally.

Then, the subproblem (50) can be seen as a particular case of the perturbed SQP subproblem (31), where for $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ and $\xi \in \mathbb{R}^n$ one takes

$$\begin{aligned} \psi((x, \lambda, \mu), \xi) &= \psi((x, \lambda), \xi) \\ &= L(x + \xi, \lambda) - \langle \lambda, h'(\bar{\pi}(x))\xi \rangle \\ &\quad - f(x) - \langle f'(x), \xi \rangle - \frac{1}{2} \left\langle \frac{\partial^2 L}{\partial x^2}(x, \lambda)\xi, \xi \right\rangle \end{aligned}$$

$$\begin{aligned}
 &= L(x + \xi, \lambda) - \left\langle \frac{\partial L}{\partial x}(x, \lambda), \xi \right\rangle - \frac{1}{2} \left\langle \frac{\partial^2 L}{\partial x^2}(x, \lambda) \xi, \xi \right\rangle \\
 &\quad - \langle \lambda, (h'(\bar{\pi}(x)) - h'(x)) \xi \rangle - f(x), \tag{51}
 \end{aligned}$$

$$\omega_2((x, \lambda, \mu), \xi) = \omega_2(x, \xi) = h'(\bar{\pi}(x))(x + \xi - \bar{\pi}(x)) - h(x) - h'(x)\xi, \tag{52}$$

$$\omega_3((x, \lambda, \mu), \xi) = 0. \tag{53}$$

If f and h are smooth enough, then it can be seen that the function ψ and the mappings ω_2 and ω_3 , defined by (51)–(53), satisfy all the requirements in Theorem 5.2, with $\chi_1(\cdot) \equiv 0$ and $\chi_2(\cdot) \equiv 0$. We thus obtain local superlinear convergence of the exact restoration scheme, again as a direct consequence of our general principles.

Theorem 5.6 *Under the hypotheses of Theorem 5.3, assume that, if $x^k = x^j$ for any two iteration indices k and j , then the feasibility phase of the exact restoration method computes $\pi^k = \pi^j$.*

Then, there exists a constant $\delta > 0$ such that, for any starting point $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, there exists a sequence $\{(x^k, \lambda^k, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ such that, for each k , the point x^{k+1} is a stationary point of problem (49), where π^k computed at the feasibility phase, and $(\lambda^{k+1} - \lambda^k, \mu^{k+1})$ is an associated Lagrange multiplier, satisfying the localization condition (30); any such sequence converges to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, and the rate of convergence is superlinear.

Of course, the exact restoration scheme is just a conceptual motivation here, as solving the subproblems (48) and (49) exactly is impractical (in most cases, simply impossible). However, once convergence of this conceptual scheme is resolved, we can naturally pass to the practical question of what kind of inexactness can be allowed when solving the subproblems involved, while still maintaining the local superlinear convergence. To that end, consider the following framework, which we refer to as the *inexact restoration method*. At each iteration, the feasibility phase now consists of computing some/any π^k satisfying

$$\|h(\pi)\| \leq \varphi_0(\|h(x^k)\|), \quad \pi \geq 0, \tag{54}$$

and the optimality phase consists of computing x^{k+1} and (η^k, μ^{k+1}) satisfying

$$\left\| \frac{\partial L}{\partial x}(x, \lambda^k) + (h'(\pi^k))^T \eta - \mu \right\| \leq \varphi_1 \left(\left\| \frac{\partial L}{\partial x}(\pi^k, \lambda^k) - \mu^k \right\| \right), \tag{55}$$

$$\|h'(\pi^k)(x - \pi^k)\| \leq \varphi_2 \left(\left\| \frac{\partial L}{\partial x}(\pi^k, \lambda^k) - \mu^k \right\| \right), \tag{56}$$

$$\mu \geq 0, \quad x \geq 0, \quad \langle \mu, x \rangle = 0, \tag{57}$$

where $\varphi_0, \varphi_1, \varphi_2 : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ are some forcing functions; the dual iterate is then given by $\lambda^{k+1} = \lambda^k + \eta^k$.

We further assume that π^k computed at the feasibility phase is within a controllable distance from x^k . Specifically, in addition to (54), it must satisfy the localization condition

$$\|\pi - x^k\| \leq K \|(x^k - \bar{x}, \lambda^k - \bar{\lambda}, \mu^k - \bar{\mu})\| \tag{58}$$

for some $K > 0$ independent of (x^k, λ^k, μ^k) . In practice, this can be achieved, e.g., by approximately solving the subproblem (48), or by other feasibility restoration strategies; see [20].

Employing (54)–(57), by the previous discussion of the exact restoration scheme, an iteration of the inexact restoration method can be interpreted in the perturbed SQP framework as (35) with ω_1 defined by (32), (33), where ψ is given by (51), with ω_2 and ω_3 defined by (52), (53), and with the appropriate χ_1 and χ_2 . Applying Theorem 5.2 again, we readily obtain local convergence properties of inexact restoration.

Theorem 5.7 *Under the hypotheses of Theorem 5.3, assume the convention that if $(x^k, \lambda^k, \mu^k) = (x^j, \lambda^j, \mu^j)$ for any two iteration indices k and j ; then, the feasibility phase of the inexact restoration method computes $\pi^k = \pi^j$.*

Then, for any functions $\varphi_0, \varphi_1, \varphi_2 : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $\varphi_0(t) = o(t)$, $\varphi_1(t) = o(t)$ and $\varphi_2(t) = o(t)$ as $t \rightarrow 0$, and any fixed $K \geq 1$, there exists $\delta > 0$ such that, for any $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, there exists an iterative sequence $\{(x^k, \pi^k, \lambda^k, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^n$ such that, for each k , the triple $(x^{k+1}, \lambda^{k+1} - \lambda^k, \mu^{k+1})$ satisfies (55)–(57), where π^k is computed at the feasibility phase and satisfies (58), and the localization condition (30) holds; for any such sequence, $\{(x^k, \lambda^k, \mu^k)\}$ converges to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, and the rate of convergence is superlinear.

5.2.4 Sequential Quadratically Constrained Quadratic Programming

The sequential quadratically constrained quadratic programming (SQCQP) method [23–28] uses a more direct way of passing the second-order information about the constraints to the subproblem, which is different from that of SQP. Specifically, both the objective function and the constraints are approximated up to second order. For the general optimization problem (7), next primal iterate x^{k+1} of the SQCQP method is defined by solving

$$\begin{aligned} &\text{minimize } f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{2} f''(x^k)[x - x^k, x - x^k] \\ \text{s.t. } & \quad h(x^k) + h'(x^k)(x - x^k) + \frac{1}{2} h''(x^k)[x - x^k, x - x^k] = 0, \\ & \quad g(x^k) + g'(x^k)(x - x^k) + \frac{1}{2} g''(x^k)[x - x^k, x - x^k] \leq 0. \end{aligned} \tag{59}$$

Observe that SQCQP is in principle a primal algorithm: dual variables are not used to formulate the subproblems.

The subproblem (59) of the SQCQP method has a quadratic objective function and *quadratic constraints*, and is generally more difficult to solve than the SQP subproblem. However, modern computational tools sometimes make solving such subproblems practical (e.g., by interior-point methods for second-order cone programming). One possible advantage of SQCQP compared to SQP is that in globalization by nonsmooth penalty function, the former allows for the unit stepsize close to the solution (under certain reasonable assumptions) without any special modifications to the algorithm [27]. In other words, Maratos [76] effect does not occur. In the case of SQP, some modifications (e.g., second-order corrections [31]) are required to prevent Maratos effect, in general.

The subproblem (59) can be seen as a particular case of the perturbed SQP subproblem (31), where for $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ and $\xi \in \mathbb{R}^n$, one takes

$$\psi((x, \lambda, \mu), \xi) = -\frac{1}{2}\langle \lambda, h''(x)[\xi, \xi] \rangle - \frac{1}{2}\langle \mu, g''(x)[\xi, \xi] \rangle, \tag{60}$$

$$\omega_2((x, \lambda, \mu), \xi) = \omega_2(x, \xi) = \frac{1}{2}h''(x)[\xi, \xi], \tag{61}$$

$$\omega_3((x, \lambda, \mu), \xi) = \omega_3(x, \xi) = \frac{1}{2}g''(x)[\xi, \xi]. \tag{62}$$

It can be verified directly that under appropriate smoothness assumptions on f, h and g , the function ψ and the mappings ω_2 and ω_3 , defined by (60)–(62), possess all the properties required in Theorem 5.2, whatever is taken as χ_1 and χ_2 . Applying Theorem 5.2 with $\chi_1(\cdot) \equiv 0$ and $\chi_2(\cdot) \equiv 0$, we readily obtain the following result on local convergence of the SQCQP method.

Theorem 5.8 *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}, h : \mathbb{R}^n \rightarrow \mathbb{R}^l$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be twice differentiable in a neighborhood of $\bar{x} \in \mathbb{R}^n$, with their second derivatives being continuous at \bar{x} . Let \bar{x} be a local solution of problem (7), satisfying the SMFCQ and the SOSC (13) for the associated unique Lagrange multiplier $(\bar{\lambda}, \bar{\mu}) \in \mathbb{R}^l \times \mathbb{R}^m$.*

Then, there exists a constant $\delta > 0$ such that, for any starting point $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, there exists a sequence $\{(x^k, \lambda^k, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ such that, for each k, x^{k+1} is a stationary point of problem (59) and $(\lambda^{k+1}, \mu^{k+1})$ is an associated Lagrange multiplier, satisfying the localization condition (30); any such sequence converges to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, and the rate of convergence is superlinear.

5.3 Stabilized Sequential Quadratic Programming

Stabilized SQP method was introduced in [11] as a tool for preserving fast convergence on problems with nonunique Lagrange multipliers associated to the solution, i.e., *degenerate* problems. The method was originally stated in the form of solving the min-max subproblems of the form

$$\begin{aligned}
 &\text{minimize} && \max_{(\lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}_+^m} \left(\langle f'(x^k), x - x^k \rangle \right. \\
 & && + \frac{1}{2} \left\langle \frac{\partial^2 L}{\partial x^2}(x^k, \lambda^k, \mu^k)(x - x^k), x - x^k \right\rangle \\
 & && + \langle \lambda, h(x^k) + h'(x^k)(x - x^k) \rangle + \langle \mu, g(x^k) + g'(x^k)(x - x^k) \rangle \\
 & && \left. - \frac{\rho_k}{2} (\|\lambda - \lambda^k\|^2 + \|\mu - \mu^k\|^2) \right) \\
 &\text{s.t.} && x \in \mathbb{R}^n,
 \end{aligned}$$

where $\rho_k > 0$ is the dual stabilization parameter. It turns out [77] that this min-max problem is equivalent to the following QP in the primal-dual space:

$$\begin{aligned}
 &\text{minimize} && \langle f'(x^k), x - x^k \rangle + \frac{1}{2} \left\langle \frac{\partial^2 L}{\partial x^2}(x^k, \lambda^k, \mu^k)(x - x^k), x - x^k \right\rangle \\
 & && + \frac{\rho_k}{2} (\|\lambda\|^2 + \|\mu\|^2) \\
 &\text{s.t.} && h(x^k) + h'(x^k)(x - x^k) - \rho_k(\lambda - \lambda^k) = 0, \\
 & && g(x^k) + g'(x^k)(x - x^k) - \rho_k(\mu - \mu^k) \leq 0.
 \end{aligned} \tag{63}$$

Note that for $\rho_k = 0$, the subproblem (63) formally becomes the usual SQP subproblem. On the other hand, for $\rho_k > 0$, the constraints in (63) have the so-called “elastic mode” feature: they are automatically consistent regardless of any constraint qualification or convexity assumptions. This is one major difference from standard SQP methods, which in particular is relevant for dealing with degenerate problems (as is well known, in the degenerate case usual SQP subproblems can simply be infeasible, and thus the method be not even well defined).

Various local convergence results for the stabilized SQP method were derived in [4, 11, 12, 78, 79]. The sharpest results were obtained in [13] for general problems assuming the SOSC (13) only, and in [14] for the equality-constrained case, assuming the weaker noncriticality property of the Lagrange multiplier (recall Remark 3.3). We emphasize again that no constraint qualifications are assumed, and thus the set of multipliers associated to a solution need not be a singleton, and can even be unbounded.

Writing the KKT system for the stabilized SQP subproblem (63) yields the following relations:

$$\begin{aligned}
 &f'(x^k) + \frac{\partial^2 L}{\partial x^2}(x^k, \lambda^k, \mu^k)(x - x^k) + (h'(x^k))^T \lambda + (g'(x^k))^T \mu = 0, \\
 & \quad h(x^k) + h'(x^k)(x - x^k) - \rho_k(\lambda - \lambda^k) = 0, \\
 & \quad \mu \geq 0, \quad g(x^k) + g'(x^k)(x - x^k) - \rho_k(\mu - \mu^k) \leq 0, \\
 & \quad \langle \mu, g(x^k) + g'(x^k)(x - x^k) - \rho_k(\mu - \mu^k) \rangle = 0.
 \end{aligned} \tag{64}$$

Let the stabilization parameter be defined as a function of the current primal-dual iterate (x^k, λ^k, μ^k) only, that is, it is given by $\rho_k = \rho(x^k, \lambda^k, \mu^k)$ for some fixed $\rho : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ (this is indeed the case for the natural choices, specified further below). Then, (64) can be interpreted as subproblem (29) of the perturbed Josephy–Newton method, where Φ and N are defined according to (9) and (10), respectively,

and where for $u = (x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ and $v = (\xi, \eta, \zeta) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ we set

$$\begin{aligned} \Omega(u, v) &= \{\omega(u, v)\}, \\ \omega(u, v) &= (0, -\rho(x, \lambda, \mu)\eta, -\rho(x, \lambda, \mu)\zeta). \end{aligned}$$

The analysis in [13] essentially entails showing that, assuming the SOSC (13), all the assumptions of Theorem 4.3 are satisfied for the specified instance of the perturbed Josephy–Newton method, with ρ defined as the residual of the KKT system (8):

$$\rho(x, \lambda, \mu) = \left\| \left(\frac{\partial L}{\partial x}(x, \lambda, \mu), h(x), \min\{\mu, -g(x)\} \right) \right\| \tag{65}$$

(to be specific, we consider the so-called natural residual given by (65), though other KKT residuals can also be used here).

Theorem 5.9 *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $h : \mathbb{R}^n \rightarrow \mathbb{R}^l$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be twice differentiable in a neighborhood of $\bar{x} \in \mathbb{R}^n$, with their second derivatives being continuous at \bar{x} . Let \bar{x} be a stationary point of problem (7), satisfying the SOSC (13) for some associated Lagrange multiplier $(\bar{\lambda}, \bar{\mu}) \in \mathbb{R}^l \times \mathbb{R}^m$. Let $\rho : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ be defined according to (65).*

Then, for any $\sigma > 0$ large enough and any $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}_+^m$ close enough to the solution point $(\bar{x}, \bar{\lambda}, \bar{\mu})$, there exists an iterative sequence $\{(x^k, \lambda^k, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ such that, for each k , $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$ satisfies (64) with $\rho_k = \rho(x^k, \lambda^k, \mu^k)$, and also satisfies the localization condition

$$\|(x^{k+1} - x^k, \lambda^{k+1} - \lambda^k, \mu^{k+1} - \mu^k)\| \leq \sigma \operatorname{dist}((x^k, \lambda^k, \mu^k), \{\bar{x}\} \times \mathcal{M}(\bar{x}));$$

any such sequence converges to $(\bar{x}, \lambda^, \mu^*)$ with some $(\lambda^*, \mu^*) \in \mathcal{M}(\bar{x})$, and the rates of convergence of the sequences $\{(x^k, \lambda^k, \mu^k)\}$ to $(\bar{x}, \lambda^*, \mu^*)$ and of $\{\operatorname{dist}((x^k, \lambda^k, \mu^k), \{\bar{x}\} \times \mathcal{M}(\bar{x}))\}$ to zero are superlinear.*

6 Beyond Perturbed Josephy–Newton Framework: Augmented Lagrangian Algorithm (Method of Multipliers)

In the previous sections, we strived to exhibit that local convergence analyses of a good variety of optimization algorithms can be conveniently cast within the perturbed Josephy–Newton framework (and the associated perturbed SQP). It is completely natural, however, that this might not be (and is not) possible for every algorithm with fast convergence, and within the framework *precisely* as stated above. However, the ideas do extend, with appropriate modifications depending on the method and problem at hand. Two reasons why the stated perturbed framework may not be applicable directly to some other “potential candidate” algorithms are the following: algorithm parameters that are not defined by smooth functions of the problem variables, and restricted smoothness of the problem data. Indeed, despite the clear generality of the

framework, there are also some restrictions, in particular with respect to the two issues just mentioned: there are no parameters in the (perturbed) Josephy–Newton scheme of Sect. 4, and it subsumes differentiability of Φ (or at least semismoothness of Φ ; though as discussed above, possible extensions in this direction are somewhat limited). At the same time, Theorems 3.1–3.3 on abstract Newtonian schemes allow for parametric methods, and do not assume any smoothness. Hence, these can still be applicable when Theorems 4.2 and 4.3 are not.

As an example of such a case, and following the approaches in [1,43,59], we shall consider the augmented Lagrangian algorithm for the optimization problem (7). This algorithm, known also as the method of multipliers, dates back to [39,40] and is one of the fundamental techniques in optimization; some other key references are [41,42,67].

Recall that, given a penalty parameter value $c > 0$, the augmented Lagrangian $L_c : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}$ for the problem (7) is defined by

$$L_c(x, \lambda, \mu) := f(x) + \frac{1}{2c} (\|\lambda + ch(x)\|^2 + \|\max\{0, \mu + cg(x)\}\|^2).$$

For some current approximations (λ^k, μ^k) of Lagrange multipliers and some $c_k > 0$, the *augmented Lagrangian method* solves the unconstrained optimization problem

$$\text{minimize } L_{c_k}(x, \lambda^k, \mu^k) \text{ s.t. } x \in \mathbb{R}^n, \tag{66}$$

and updates the dual iterates by

$$\lambda^{k+1} = \lambda^k + c_k h(x^{k+1}), \quad \mu^{k+1} = \max\{0, \mu^k + c_k g(x^{k+1})\}. \tag{67}$$

As solving (66) to optimality is generally impractical, an approximate stationary point is usually computed, in the sense that x^{k+1} is required to satisfy the condition

$$\left\| \frac{\partial L_{c_k}}{\partial x}(x, \lambda^k, \mu^k) \right\| \leq \tau_k, \tag{68}$$

for some tolerance parameter $\tau_k \geq 0$. Then, $(\lambda^{k+1}, \mu^{k+1})$ is computed by (67).

Iterations of this method use the parameter c_k given by certain update rules, usually not depending continuously on the iterates. Therefore, the method does not fit the Josephy–Newton framework. Note also that the iterations do not employ any kind of linearizations or other approximations of the problem data; in particular, they do not involve second derivatives of f , h , and g (and hence, the first derivative of the corresponding GE mapping Φ). Nevertheless, as explicitly shown in [1] (and somewhat implicitly in [43]), the method can be analyzed by the abstract Newtonian frameworks of Sect. 3, leading to local convergence properties that are stronger than previously available. In this sense, quite surprisingly, the augmented Lagrangian method can also be viewed as of Newtonian type. The sharpest known local convergence results for this method were obtained in [43] for the case of twice differentiable problem data, and in [1] for the case when f , h , and g are differentiable near the solution \bar{x} , and their derivatives are locally Lipschitz-continuous at \bar{x} . Here, we adopt the latter setting.

By (67) and a direct computation, it can be seen that

$$\frac{\partial L_{c_k}}{\partial x}(x^{k+1}, \lambda^k, \mu^k) = \frac{\partial L}{\partial x}(x^{k+1}, \lambda^{k+1}, \mu^{k+1}),$$

$$h(x^{k+1}) - \frac{1}{c_k}(\lambda^{k+1} - \lambda^k) = 0, \quad \min\{\mu^{k+1}, -c_k g(x^{k+1}) + (\mu^{k+1} - \mu^k)\} = 0.$$

Suppose that the tolerance parameter τ_k is chosen as a function of the current iterate, i.e., $\tau_k = \tau(x^k, \lambda^k, \mu^k)$ with some function $\tau : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ (for this parameter, choices of this type make good sense; one reasonable option would be to take τ_k as the natural residual (65) of the KKT system (8) of the problem (7)). Then, the iteration of the (inexact) augmented Lagrangian method, given by (68), (67), can be written in the form of the GE (15) with $\Pi = \mathbb{R}_+ \setminus \{0\}$, and the multifunction \mathcal{A} defined by

$$\mathcal{A}(c, \tilde{u}, u) := \begin{pmatrix} \frac{\partial L}{\partial x}(x, \lambda, \mu) + B(0, \tau(\tilde{x}, \tilde{\lambda}, \tilde{\mu})) \\ h(x) - \frac{1}{c}(\lambda - \tilde{\lambda}) \\ -g(x) + \frac{1}{c}(\mu - \tilde{\mu}) \end{pmatrix}, \tag{69}$$

where $\tilde{u} = (\tilde{x}, \tilde{\lambda}, \tilde{\mu})$, $u = (x, \lambda, \mu)$, and N is given by (10).

Suppose first that $\tau(\cdot) \equiv 0$, i.e., consider that the method computes stationary points of subproblems exactly. Then, it can be readily seen that

$$\mathcal{A}(c, \tilde{u}, \tilde{u}) = \Phi(\tilde{u}), \quad \Phi(u) - \mathcal{A}(c, \tilde{u}, u) = \left(0, \frac{1}{c}(\lambda - \tilde{\lambda}), -\frac{1}{c}(\mu - \tilde{\mu})\right),$$

and hence

$$\|(\Phi(u^1) - \mathcal{A}(c, \tilde{u}, u^1)) - (\Phi(u^2) - \mathcal{A}(c, \tilde{u}, u^2))\| = \frac{1}{c}\|(\lambda^1 - \lambda^2, \mu^1 - \mu^2)\|.$$

This implies that (20) holds with $\omega(c, \tilde{u}, u^1, u^2) = 1/c$ and $u^1 = (x^1, \lambda^1, \mu^1)$, $u^2 = (x^2, \lambda^2, \mu^2)$ for any $c > 0$ and any $x^1, x^2 \in \mathbb{R}^n$, and any $\lambda^1, \lambda^2 \in \mathbb{R}^l$, $\mu^1, \mu^2 \in \mathbb{R}^m$. Therefore, if we take $\Pi = [\bar{c}, +\infty[$ for a sufficiently large $\bar{c} > 0$, then the local convergence and rate of convergence result for the exact method of multipliers follows readily from Theorem 3.1 under the LICQ and the SSOSC appropriately extended to the reduced smoothness setting at hand:

$$\forall H \in \partial_x \frac{\partial L}{\partial x}(\bar{x}, \bar{\lambda}, \bar{\mu}) \quad \langle H\xi, \xi \rangle > 0, \quad \forall \xi \in C_+(\bar{x}, \bar{\mu}) \setminus \{0\}. \tag{70}$$

This is because, according to the results in [54,59], strong metric regularity of the GE solution $\bar{u} = (\bar{x}, \bar{\lambda}, \bar{\mu})$ is implied by the combination of the LICQ and of the SSOSC (70).

Theorem 6.1 *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $h : \mathbb{R}^n \rightarrow \mathbb{R}^l$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be differentiable in a neighborhood of $\bar{x} \in \mathbb{R}^n$, with their derivatives being locally Lipschitz-continuous at \bar{x} . Let \bar{x} be a stationary point of problem (7), satisfying the LICQ and the SSOSC (70) for the associated unique Lagrange multiplier $(\bar{\lambda}, \bar{\mu}) \in \mathbb{R}^l \times \mathbb{R}^m$.*

Then, there exist constants $\bar{c} > 0$ and $\delta > 0$ such that, for any starting point $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, and any sequence $\{c_k\} \subset [\bar{c}, +\infty[$, there exists a unique sequence $\{(x^k, \lambda^k, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ such that, for all k , it holds that x^{k+1} is a stationary point of problem (66), the pair $(\lambda^{k+1}, \mu^{k+1})$ satisfies (67), and the localization condition (30) holds; this sequence converges to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, and the rate of convergence is linear. Moreover, the rate of convergence is superlinear if $c_k \rightarrow +\infty$.

Suppose now that the subproblems minimization tolerance τ satisfies

$$\tau(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \leq \theta(\tilde{x}, \tilde{\lambda}, \tilde{\mu})\|(\tilde{x} - \bar{x}, \tilde{\lambda} - \bar{\lambda}, \tilde{\mu} - \bar{\mu})\| \quad \forall (\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m,$$

where $\theta : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ is a function such that $\theta(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \rightarrow 0$ as $(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \rightarrow (\bar{x}, \bar{\lambda}, \bar{\mu})$. In this case, for \mathcal{A} defined in (69), assumption (ii) of Theorem 3.2 holds with $\Pi = [\bar{c}, +\infty[$ for a sufficiently large $\bar{c} > 0$, and with $\omega(c, \tilde{u}, u) = \max\{1/c, \theta(\tilde{x}, \tilde{\lambda}, \tilde{\mu})\}$. Constructive and practically relevant choices of a function τ with the needed properties can be based on residuals of the KKT system (8). For instance, one can take any τ such that

$$\tau(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) = o(\rho(\tilde{x}, \tilde{\lambda}, \tilde{\mu})), \tag{71}$$

where ρ is the KKT natural residual, defined in (65).

Verifying the assumptions (i) and (iii) of Theorem 3.2 is a more subtle issue, but it can be seen that both are implied by the combination of the SMFCQ and of the SOSC of the form

$$\forall H \in \partial_x \frac{\partial L}{\partial x}(\bar{x}, \bar{\lambda}, \bar{\mu}) \quad \langle H\xi, \xi \rangle > 0, \quad \forall \xi \in C(\bar{x}) \setminus \{0\}. \tag{72}$$

Theorem 6.2 *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $h : \mathbb{R}^n \rightarrow \mathbb{R}^l$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be differentiable in a neighborhood of $\bar{x} \in \mathbb{R}^n$, with their derivatives being locally Lipschitz-continuous at \bar{x} . Let \bar{x} be a stationary point of problem (7), satisfying the SMFCQ and the SOSC (72) for the associated unique Lagrange multiplier $(\bar{\lambda}, \bar{\mu}) \in \mathbb{R}^l \times \mathbb{R}^m$. Let $\tau : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ be a function satisfying $\tau(x, \lambda, \mu) = o(\|(x - \bar{x}, \lambda - \bar{\lambda}, \mu - \bar{\mu})\|)$.*

Then, there exist constants $\bar{c} > 0$ and $\delta > 0$ such that, for any starting point $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, and any sequence $\{c_k\} \subset [\bar{c}, +\infty[$, there exists a sequence $\{(x^k, \lambda^k, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ such that, for all k , the point x^{k+1} satisfies (68) with $\tau_k = \tau(x^k, \lambda^k, \mu^k)$, and (67) and the localization condition (30) hold; any such sequence converges to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, and the rate of convergence is linear. Moreover, the rate of convergence is superlinear if $c_k \rightarrow +\infty$.

Consider now a stationary point $\bar{x} \in \mathbb{R}^n$ of the optimization problem (7), and some particular Lagrange multiplier $(\bar{\lambda}, \bar{\mu}) \in \mathcal{M}(\bar{x})$ which need no longer be unique (i.e.,

the problem can be degenerate). It was established in [58, Corollary 2.1] that in this setting, assumption (i) of Theorem 3.3 with $\bar{u} = (\bar{x}, \bar{\lambda}, \bar{\mu})$ and with \bar{U} replaced by its subset $\{\bar{x}\} \times \mathcal{M}(\bar{x})$ is implied by the SOSC (72).

Suppose further that a function τ satisfies the condition

$$\begin{aligned} \tau(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) &\leq \theta(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \operatorname{dist}((\tilde{x}, \tilde{\lambda}, \tilde{\mu}), \{\bar{x}\} \times \mathcal{M}(\bar{x})) \\ \forall (\tilde{x}, \tilde{\lambda}, \tilde{\mu}) &\in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m, \end{aligned}$$

where $\theta : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ is again a function such that $\theta(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \rightarrow 0$ as $(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \rightarrow (\bar{x}, \bar{\lambda}, \bar{\mu})$. Then, for any $\sigma > 0$, assumption (ii) of Theorem 3.3 holds for the inexact augmented Lagrangian method with $\Pi = [\bar{c}, +\infty[$ and $\omega(c, \tilde{u}, u) = \sigma/c + \theta(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$ for any sufficiently large $\bar{c} > 0$. Note that (71) still gives a relevant practical rule for choosing τ with the needed properties.

Finally, assumption (iii) of Theorem 3.3 (with Π defined above) can be verified for any $\sigma > 0$ and any sufficiently large $\bar{c} > 0$, employing the results in [43].

Theorem 6.3 *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $h : \mathbb{R}^n \rightarrow \mathbb{R}^l$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be differentiable in a neighborhood of a point $\bar{x} \in \mathbb{R}^n$, with their derivatives being locally Lipschitz-continuous at \bar{x} . Let \bar{x} be a stationary point of the problem (7), satisfying the SOSC (72) for an associated Lagrange multiplier $(\bar{\lambda}, \bar{\mu}) \in \mathbb{R}^l \times \mathbb{R}^m$. Let $\tau : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ be a function satisfying $\tau(x, \lambda, \mu) = o(\operatorname{dist}((x, \lambda, \mu), \{\bar{x}\} \times \mathcal{M}(\bar{x})))$.*

Then, for any $\sigma > 0$, there exists $\bar{c} > 0$ such that, for any starting point $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ close enough to $(\bar{x}, \bar{\lambda}, \bar{\mu})$, and any sequence $\{c_k\} \subset [\bar{c}, +\infty[$, there exists a sequence $\{(x^k, \lambda^k, \mu^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m$ such that, for all k , the point x^{k+1} satisfies (68) with $\tau_k = \tau(x^k, \lambda^k, \mu^k)$, and (67) and the localization condition

$$\|(x^{k+1} - x^k, \lambda^{k+1} - \lambda^k, \mu^{k+1} - \mu^k)\| \leq \sigma (\|x^k - \bar{x}\| + \operatorname{dist}((\lambda^k, \mu^k), \mathcal{M}(\bar{x})))$$

hold; any such sequence converges to $(\bar{x}, \lambda^, \mu^*)$ with some $(\lambda^*, \mu^*) \in \mathcal{M}(\bar{x})$, and the rates of convergence of the sequences $\{(x^k, \lambda^k, \mu^k)\}$ to $(\bar{x}, \lambda^*, \mu^*)$ and of $\{\|x^k - \bar{x}\| + \operatorname{dist}((\lambda^k, \mu^k), \mathcal{M}(\bar{x}))\}$ to zero are linear. Moreover, both of those rates are superlinear if $c_k \rightarrow +\infty$.*

For the case of twice differentiable problem data, Theorem 6.3 was obtained in [43]. Note that previous literature on augmented Lagrangian methods required, in addition to the SOSC, the LICQ, and the strict complementarity condition (or the LICQ and the SSOSC).

The comments above concerned with the role and nature of localization conditions remain valid for augmented Lagrangian methods. Moreover, at least at first glance, from the practical viewpoint these conditions may look more problematic here: when solving the unconstrained minimization subproblem (66), one seeks a point with the smallest possible value of the augmented Lagrangian (regardless of this point being close to the previous primal iterate or not). However, the localization condition is again not unreasonable because of the following. In the process of proving that the minimization subproblems (66) have solutions, it is established that the augmented

Lagrangian is locally coercive (has uniform local quadratic growth) [43, Prop. 3.1]. Thus, if a typical unconstrained minimization method for solving (66) will use a starting point in this region (e.g., the previous primal iterate), then the process should be expected to converge to a minimizer in this region (in fact, at a fast rate) and not “jump far away,” even if the augmented Lagrangian has far away minimizers with lower values.

We finally comment that similarly to the method of multipliers, the LCL method considered in Sect. 5.2.2 also does not require second derivatives of the problem data, but in the absence of twice differentiability, it does not fit the perturbed semismooth Josephy–Newton framework. Yet, again as for the method of multipliers above, LCL local convergence can be fully analyzed by the abstract Newtonian frameworks of Sect. 3; see [1].

7 Conclusions

We have presented a survey of state-of-the-art in local convergence theories for a wide class of numerical methods for variational and optimization problems. We believe that methods of this class can naturally be regarded as Newtonian, if a reasonably broad understanding is adopted of what is a Newtonian family. A number of issues related to theoretical results discussed in this survey are waiting for further development. Some directions of future research are outlined above: among them are relaxations of the smoothness and/or regularity assumptions, and globalization strategies combining the attractive local convergence properties of the algorithms with guaranties of reasonable global behavior. Contributions in those directions are currently the subject of interest for many research groups throughout the world.

Acknowledgments Research of the first author is supported by the Russian Foundation for Basic Research Grant 14-01-00113. The second author is supported in part by CNPq Grant 302637/2011-7, by PRONEX–Optimization, and by FAPERJ.

References

1. Izmailov, A.F., Kurennoy, A.S.: Abstract Newtonian frameworks and their applications. *SIAM J. Optim.* **23**, 2369–2396 (2013)
2. Facchinei, F., Pang, J.-S.: *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer-Verlag, New York (2003)
3. Bonnans, J.F.: Local analysis of Newton-type methods for variational inequalities and nonlinear programming. *Appl. Math. Optim.* **29**, 161–186 (1994)
4. Fischer, A.: Local behavior of an iterative framework for generalized equations with nonisolated solutions. *Math. Program.* **94**, 91–124 (2002)
5. Klatte, D., Kummer, B.: *Nonsmooth Equations in Optimization: Regularity, Calculus, Methods and Applications*. Kluwer Academic Publishers, Dordrecht (2002)
6. Robinson, S.M.: Newton’s method for a class of nonsmooth functions. *Set-Valued Anal.* **2**, 291–305 (1994)
7. Robinson, S.M.: A point-of-attraction result for Newton’s method with point-based approximations. *Optimization* **60**, 89–99 (2011)
8. Dontchev, A.L., Rockafellar, R.T.: *Implicit Functions and Solution Mappings*. Springer, New York (2009)

9. Izmailov, A.F., Solodov, M.V.: Inexact Josephy–Newton framework for generalized equations and its applications to local analysis of Newtonian methods for constrained optimization. *Comput. Optim. Appl.* **46**, 347–368 (2010)
10. Izmailov, A.F., Solodov, M.V.: A truncated SQP method based on inexact interior-point solutions of subproblems. *SIAM J. Optim.* **20**, 2584–2613 (2010)
11. Wright, S.J.: Superlinear convergence of a stabilized SQP method to a degenerate solution. *Comput. Optim. Appl.* **11**, 253–275 (1998)
12. Hager, W.W.: Stabilized sequential quadratic programming. *Comput. Optimizat. Appl.* **12**, 253–273 (1999)
13. Fernández, D., Solodov, M.: Stabilized sequential quadratic programming for optimization and a stabilized Newton-type method for variational problems. *Math. Program.* **125**, 47–73 (2010)
14. Izmailov, A.F., Solodov, M.V.: Stabilized SQP revisited. *Math. Program.* **133**, 93–120 (2012)
15. Solodov, M.V.: Constraint qualifications. In: Cochran, J.J. (ed.) *Wiley Encyclopedia of Operations Research and Management Science*. Wiley, New York (2010)
16. Robinson, S.M.: A quadratically convergent algorithm for general nonlinear programming problems. *Math. Program.* **3**, 145–156 (1972)
17. Murtagh, B.A., Saunders, M.A.: A projected Lagrangian algorithm and its implementation for sparse nonlinear constraints. *Math. Program. Study* **16**, 84–117 (1982)
18. Friedlander, M.P., Saunders, M.A.: A globally convergent linearly constrained Lagrangian method for nonlinear optimization. *SIAM J. Optim.* **15**, 863–897 (2005)
19. Martínez, J.M.: Inexact restoration method with Lagrangian tangent decrease and new merit function for nonlinear programming. *J. Optim. Theory Appl.* **111**, 39–58 (2001)
20. Birgin, E.G., Martínez, J.M.: Local convergence of an inexact-restoration method and numerical experiments. *J. Optim. Theory Appl.* **127**, 229–247 (2005)
21. Fischer, A., Friedlander, A.: A new line search inexact restoration approach for nonlinear programming. *Comput. Optim. Appl.* **46**, 333–346 (2010)
22. Fernández, D., Pilotta, E.A., Torres, G.A.: An inexact restoration strategy for the globalization of the sSQP method. *Comput. Optim. Appl.* **54**, 595–617 (2013)
23. Wiest, E.J., Polak, E.: A generalized quadratic programming-based phase-I–II method for inequality-constrained optimization. *Appl. Math. Optim.* **26**, 223–252 (1992)
24. Kruk, S., Wolkowicz, H.: Sequential, quadratically constrained, quadratic programming for general nonlinear programming. In: Wolkowicz, H., Saigal, R., Vandenberghe, L. (eds.) *Handbook of Semidefinite Programming*, pp. 563–575. Kluwer Academic Publishers, Dordrecht (2000)
25. Anitescu, M.: A superlinearly convergent sequential quadratically constrained quadratic programming algorithm for degenerate nonlinear programming. *SIAM J. Optim.* **12**, 949–978 (2002)
26. Fukushima, M., Luo, Z.-Q., Tseng, P.: A sequential quadratically constrained quadratic programming method for differentiable convex minimization. *SIAM J. Optim.* **13**, 1098–1119 (2003)
27. Solodov, M.V.: On the sequential quadratically constrained quadratic programming methods. *Math. Oper. Res.* **29**, 64–79 (2004)
28. Fernández, D., Solodov, M.V.: On local convergence of sequential quadratically-constrained quadratic-programming type methods, with an extension to variational problems. *Comput. Optim. Appl.* **39**, 143–160 (2008)
29. Vardi, A.: A trust region algorithm for equality constrained minimization: convergence properties and implementation. *SIAM J. Numer. Anal.* **22**, 575–591 (1985)
30. Omojokun, E.O.: Trust region algorithms for optimization with nonlinear equality and inequality constraints. Ph.D. thesis. Department of Computer Science, University of Colorado at Boulder, (1989)
31. Fletcher, R.: Second order corrections for non-differentiable optimization. In: Griffiths, D. (ed.) *Numerical Analysis*, pp. 85–114. Springer-Verlag, Berlin (1982)
32. Herskovits, J.: A two-stage feasible direction algorithm including variable metric techniques for nonlinear optimization problems. *Rapport de Recherche 118*. INRIA, Rocquencourt, (1982)
33. Herskovits, J.: A two-stage feasible directions algorithm for nonlinear constrained optimization. *Math. Program.* **36**, 19–38 (1986)
34. Herskovits, J.: Feasible direction interior-point technique for nonlinear optimization. *J. Optim. Theory Appl.* **99**, 121–146 (1998)
35. Panier, E.R., Tits, A.L., Herskovits, J.: A QP-free, globally convergent, locally superlinearly convergent algorithm for inequality constrained optimization. *SIAM J. Control Optim.* **26**, 788–811 (1988)

36. Tits, A.L., Wächter, A., Bakhtiari, S., Urban, T.J., Lawrence, C.T.: A primal-dual interior-point method for nonlinear programming with strong global and local convergence properties. *SIAM J. Optim.* **14**, 173–199 (2003)
37. Izmailov, A.F., Kurennoy, A.S., Solodov, M.V.: Some composite-step constrained optimization methods interpreted via the perturbed sequential quadratic programming framework. *Optim. Method. Softw.* (to appear)
38. Izmailov, A.F., Solodov, M.V.: *Newton-Type Methods for Optimization and Variational Problems*. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Switzerland (2014)
39. Hestenes, M.R.: Multiplier and gradient methods. *J. Optim. Theory Appl.* **4**, 303–320 (1969)
40. Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In: Fletcher, R. (ed.) *Optimization*, pp. 283–298. Academic Press, New York (1969)
41. Conn, A.R., Gould, N., Sartenaer, A., Toint, P.L.: Convergence properties of an augmented Lagrangian algorithm for optimization with a combination of general equality and linear constraints. *SIAM J. Optim.* **6**, 674–703 (1996)
42. Andreani, R., Birgin, E.G., Martínez, J.M., Schuverdt, M.L.: On augmented Lagrangian methods with general level constraints. *SIAM J. Optim.* **18**, 1286–1309 (2007)
43. Fernández, D., Solodov, M.V.: Local convergence of exact and inexact augmented Lagrangian methods under the second-order sufficient optimality condition. *SIAM J. Optim.* **22**, 384–407 (2012)
44. Josephy, N.H.: Newton's method for generalized equations. Technical Summary Report 1965. Mathematics Research Center, University of Wisconsin, Madison, WI (1979)
45. Rockafellar, R.T., Wets, J.B.: *Variational Analysis*. Springer-Verlag, Berlin (1998)
46. Kinderlehrer, D., Stampacchia, G.: *An Introduction to Variational Inequalities and Their Applications*. Academic Press, New York (1980)
47. Bonnans, J.F., Shapiro, A.: *Perturbation Analysis of Optimization Problems*. Springer-Verlag, New York (2000)
48. Giannessi, F.: *Constrained Optimization and Image Space Analysis*. Springer-Verlag, New York (2005)
49. Boggs, P., Tolle, J.: Sequential quadratic programming. *Acta Numer.* **4**, 1–51 (1995)
50. Gill, P.E., Wong, E.: Sequential quadratic programming methods. In: Lee, J., Leyffer, S. (eds.) *Mixed Integer Nonlinear Programming*. The IMA Volumes in Mathematics and its Applications, vol. 154, pp. 147–224. Springer-Verlag, Berlin (2012)
51. Rademacher, H.: Über partielle und totale differenzierbarkeit I. *Math. Ann.* **89**, 340–359 (1919)
52. Dontchev, A.L., Rockafellar, R.T.: Newton's method for generalized equations: a sequential implicit function theorem. *Math. Program.* **123**, 139–159 (2010)
53. Robinson, S.M.: Strongly regular generalized equations. *Math. Oper. Res.* **5**, 43–62 (1980)
54. Izmailov, A.F.: Strongly regular nonsmooth generalized equations. *Math. Program.* (2013). doi:[10.1007/s10107-013-0717-1](https://doi.org/10.1007/s10107-013-0717-1)
55. Kojima, M.: Strongly stable stationary solutions in nonlinear programs. In: Robinson, S.M. (ed.) *Analysis and Computation of Fixed Points*, pp. 93–138. Academic Press, New York (1980)
56. Bonnans, J.F., Sulem, A.: Pseudopower expansion of solutions of generalized equations and constrained optimization. *Math. Program.* **70**, 123–148 (1995)
57. Dontchev, A.L., Rockafellar, R.T.: Characterizations of strong regularity for variational inequalities over polyhedral convex sets. *SIAM J. Optim.* **6**, 1087–1105 (1996)
58. Izmailov, A.F., Kurennoy, A.S., Solodov, M.V.: A note on upper Lipschitz stability, error bounds, and critical multipliers for Lipschitz-continuous KKT systems. *Math. Program.* **142**, 591–604 (2013)
59. Izmailov, A.F., Kurennoy, A.S., Solodov, M.V.: The Josephy–Newton method for semismooth generalized equations and semismooth SQP for optimization. *Set-Valued Var. Anal.* **21**, 17–45 (2013)
60. Josephy, N.H.: Quasi-Newton methods for generalized equations. Technical Summary Report 1966. Mathematics Research Center, University of Wisconsin, Madison, WI (1979)
61. Mifflin, R.: Semismooth and semiconvex functions in constrained optimization. *SIAM J. Control Optim.* **15**, 959–972 (1977)
62. Conn, A.R., Gould, N.I.M., Toint, PhL: *Trust-Region Methods*. SIAM, Philadelphia (2000)
63. Bonnans, J.F., Gilbert, JCh., Lemaréchal, C., Sagastizábal, C.: *Numerical Optimization: Theoretical and Practical Aspects*, 2nd edn. Springer-Verlag, Berlin (2006)
64. Nocedal, J., Wright, S.J.: *Numerical Optimization*, 2nd edn. Springer, New York (2006)
65. Wilson, R.B.: A simplicial algorithm for concave programming. Ph.D. thesis. Graduate School of Business Administration, Harvard University (1963)

66. Robinson, S.M.: Perturbed Kuhn–Tucker points and rates of convergence for a class of nonlinear-programming algorithms. *Math. Program.* **7**, 1–16 (1974)
67. Bertsekas, D.P.: *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, New York (1982)
68. Fernández, D., Izmailov, A.F., Solodov, M.V.: Sharp primal superlinear convergence results for some Newtonian methods for constrained optimization. *SIAM J. Optim.* **20**, 3312–3334 (2010)
69. Dembo, R.S., Eisenstat, S.C., Steihaug, T.: Inexact Newton methods. *SIAM J. Numer. Anal.* **19**, 400–408 (1982)
70. Gould, N.I.M.: Some reflections on the current state of active-set and interior-point methods for constrained optimization. Numerical Analysis Group Internal Report 2003–1. Computational Science and Engineering Department, Rutherford Appleton Laboratory, Oxfordshire (2003)
71. Gould, N.I.M., Orban, D., Toint, PhL: Numerical methods for large-scale nonlinear optimization. *Acta Numer.* **14**, 299–361 (2005)
72. Leibfritz, F., Sachs, E.W.: Inexact SQP interior point methods and large scale optimal control problems. *SIAM J. Control Optim.* **38**, 272–293 (1999)
73. Murtagh, B.A., Saunders, M.A.: MINOS 5.0 user’s guide. Technical Report SOL 83.20. Stanford University (1983)
74. Martínez, J.M., Pilotta, E.A.: Inexact restoration algorithms for constrained optimization. *J. Optim. Theory Appl.* **104**, 135–163 (2000)
75. Martínez, J.M., Pilotta, E.A.: Inexact restoration methods for nonlinear programming: advances and perspectives. In: Qi, L.Q., Teo, K.L., Yang, X.Q. (eds.) *Optimization and Control with Applications*, pp. 271–292. Springer, Berlin (2005)
76. Maratos, N.: Exact penalty function algorithms for finite dimensional and control optimization problems. Ph.D. thesis. University of London (1978)
77. Li, D.-H., Qi, L.: Stabilized SQP method via linear equations. Applied Mathematics Technical Reptort AMR00/5. University of New South Wales, Sydney (2000)
78. Wright, S.J.: Modifying SQP for degenerate problems *SIAM. J. Optim.* **13**, 470–497 (2002)
79. Wright, S.J.: Constraint identification and algorithm stabilization for degenerate nonlinear programs. *Math. Program.* **95**, 137–160 (2003)