Mathematical Programming (2020) 182:37–55
https://doi.org/10.1007/s10107-019-01384-1

FULL LENGTH PAPER

Series A

On the cost of solving augmented Lagrangian subproblems

Damián Fernández[1] · Mikhail Solodov[2]

Received: 31 March 2018 / Accepted: 25 February 2019 / Published online: 8 March 2019

Abstract

At each iteration of the augmented Lagrangian algorithm, a nonlinear subproblem is being solved. The number of inner iterations (of some/any method) needed to obtain a solution of the subproblem, or even a suitable approximate stationary point, is in principle unknown. In this paper we show that to compute an approximate stationary point sufficient to guarantee local superlinear convergence of the augmented Lagrangian iterations, it is enough to solve two quadratic programming problems (or two linear systems in the equality-constrained case). In other words, two inner Newtonian iterations are sufficient. To the best of our knowledge, such results are not available even under the strongest assumptions (of second-order sufficiency, strict complementarity, and the linear independence constraint qualification). Our analysis is performed under second-order sufficiency only, which is the weakest assumption for obtaining local convergence and rate of convergence of outer iterations of the augmented Lagrangian algorithm. The structure of the quadratic problems in question is related to the stabilized sequential quadratic programming and to second-order corrections.

Keywords Augmented Lagrangian · Newton methods · Stabilized sequential quadratic programming · Second-order correction · Superlinear convergence

Mathematics Subject Classification 90C30 · 90C33 · 90C55 · 65K05

Research of the second author is supported in part by CNPq Grant 303724/2015-3, by FAPERJ Grant 203.052/2016, and by Russian Foundation for basic research Grant 19-51-12003 NNIOa.

✉ Mikhail Solodov
solodov@impa.br

Damián Fernández
dfernandez@unc.edu.ar

1 CIEM-CONICET and FaMAF, Universidad Nacional de Córdoba, Medina Allende s/n, 5000 Córdoba, Argentina

2 IMPA – Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil

1 Introduction

Given twice continuously differentiable $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$, we consider the general nonlinear constrained optimization problem

$$\begin{aligned} &\min\ f(x) \\ &\text{s.t. } g_i(x) = 0, \quad i = 1, \ldots, l \\ &\qquad g_i(x) \leq 0, \quad i = l+1, \ldots, m, \end{aligned} \tag{1}$$

where $1 \leq l \leq m$. Let $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ denote the usual Lagrangian of problem (1), i.e.,

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

The Karush–Kuhn–Tucker (KKT) system characterizing stationary points and Lagrange multipliers of problem (1) is given by

$$\begin{aligned} &0 = \frac{\partial L}{\partial x}(x, \mu), \\ &0 = g_i(x), \quad i = 1, \ldots, l \\ &0 \leq \mu_i,\ g_i(x) \leq 0,\ \mu_i g_i(x) = 0, \quad i = l+1, \ldots, m. \end{aligned} \tag{2}$$

For a given stationary point $\bar{x}$ of problem (1), we denote by $\mathcal{M}(\bar{x})$ the set of the associated Lagrange multipliers, i.e., the set of μ that satisfy (2) for $x = \bar{x}$.

To measure violation of the KKT conditions (2), we shall use the natural residual function $r : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_+$, which is given by

$$r(x, \mu) = \left(\left\| \frac{\partial L}{\partial x}(x, \mu) \right\|^2 + \sum_{i=1}^{l} (g_i(x))^2 + \sum_{i=l+1}^{m} (\mu_i - \max\{0, \mu_i + g_i(x)\})^2 \right)^{1/2}. \tag{3}$$

Note that $(\bar{x}, \bar{\mu})$ is a solution of (2) if and only if $r(\bar{x}, \bar{\mu}) = 0$.

We say that the second-order sufficient optimality condition (SOSC) holds at $(\bar{x}, \bar{\mu})$, $\bar{\mu} \in \mathcal{M}(\bar{x})$, if

$$\left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\mu})u, u \right\rangle > 0 \quad \forall u \in \mathcal{C} \setminus \{0\}, \tag{4}$$

where $\mathcal{C} = \mathcal{C}(\bar{x})$ is the critical cone of problem (1) at $\bar{x}$, i.e.,

$$\mathcal{C} = \left\{ u \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \langle f'(\bar{x}), u \rangle = 0,\ \langle g_i'(\bar{x}), u \rangle = 0 \text{ for } i \in \{1, \ldots, l\}, \\ \langle g_i'(\bar{x}), u \rangle \leq 0 \text{ for } i \in \{l+1, \ldots, m\} \text{ with } g_i(\bar{x}) = 0 \end{array} \right\}. \tag{5}$$

Recall that SOSC (4) at $(\bar{x}, \bar{\mu})$ implies that the natural residual $r(x, \mu)$ provides the local (Lipschitzian) error bound [10,14,23], i.e., there exists a constant $\beta > 0$ such that

$$\|x - \bar{x}\| + \operatorname{dist}(\mu, \mathcal{M}(\bar{x})) \leq \beta\, r(x, \mu), \tag{6}$$

for all (x, μ) close enough to $(\bar{x}, \bar{\mu})$. Moreover, this error bound is equivalent to the noncriticality property of the multiplier $\bar{\mu}$ [16], [18, Proposition 1.43] (see [18, Definition 1.41] for the definition of critical/noncritical multipliers). In particular, SOSC implies noncriticality.

One of the fundamental methods for problem (1) is the augmented Lagrangian algorithm; see the classical [15,21,22], and the recent book [4] among many others. The augmented Lagrangian function $\bar{L} : \mathbb{R}^n \times \mathbb{R}^m \times (0, \infty) \to \mathbb{R}$ is given by

$$\begin{aligned}\bar{L}(x, \mu; \sigma) = f(x) + \sum_{i=1}^{l} \left(\mu_i g_i(x) + \frac{1}{2\sigma} g_i(x)^2 \right) \\ + \frac{\sigma}{2} \sum_{i=l+1}^{m} \left(\max \left\{ 0, \mu_i + \frac{1}{\sigma} g_i(x) \right\}^2 - \mu_i^2 \right),\end{aligned}$$

where $\sigma > 0$ is the penalty parameter.

Given the current primal–dual iterate $(x^k, \mu^k) \in \mathbb{R}^n \times \mathbb{R}^m$ and the penalty parameter $\sigma_k > 0$, the exact augmented Lagrangian method generates the next iterate (x^{k+1}, μ^{k+1}) according to the following scheme:

$$\begin{aligned} &x^{k+1} \text{ is a solution of } \underset{x \in \mathbb{R}^n}{\text{minimize}}\, \bar{L}(x, \mu^k; \sigma_k), \\ &\mu_i^{k+1} = \mu_i^k + \frac{1}{\sigma_k} g_i(x^{k+1}), \quad i = 1, \ldots, l, \\ &\mu_i^{k+1} = \max \left\{ 0, \mu_i^k + \frac{1}{\sigma_k} g_i(x^{k+1}) \right\}, \quad i = l+1, \ldots, m. \end{aligned} \tag{7}$$

For practical reasons, minimization of the augmented Lagrangian in the first step in (7) is performed approximately. In fact, rather than a minimizer, an approximate stationary point is computed: at the k-th iteration, x^{k+1} satisfying

$$\left\| \frac{\partial \bar{L}}{\partial x}(x^{k+1}, \mu^k; \sigma_k) \right\| \le \varepsilon_k \tag{8}$$

is obtained. For purposes of global convergence, $\{\varepsilon_k\}$ can be any exogenous sequence of scalars converging to zero; see, e.g., [1,2]. For purposes of establishing convergence rates, however, this sequence certainly cannot be arbitrary and should be appropriately controled [9].

Note that the next primal iterate x^{k+1} is obtained by applying some algorithm to the unconstrained problem in (7), until inner iterations of this algorithm generate a point satisfying the approximate stationarity condition (8). In principle, computational effort of this inner loop is unknown. To the best of our knowledge, it is unkown even locally, i.e., close to a solution of problem (1), and regardless of the assumptions imposed (at least natural assumptions not involving convexity). In this paper, we show that close to a solution of (1), solving two quadratic programs (QPs) is sufficient to minimize the augmented Lagrangian accurately enough to ensure superlinear convergence of the overall algorithm under the second-order sufficient condition only, which is the weakest assumption for convergence rates of (outer iterations of) the augmented Lagrangian algorithm, even in the case of exact solution of subproblems. To the best of our knowledge, the cost of solving those subproblems had not been determined previously, even under much stronger assumptions (such as adding to second-order sufficiency strict complementarity and the linear independece constraint qualification). Thus, we show that the augmented Lagrangian algorithm can be locally reduced to a kind of two-step Newton process, converging superlinearly. Here, it must be commented that the idea of trying to (approximately) solve the augmented Lagrangian subproblems by some Newton steps, at least asymptotically, is certainly not new. We shall mention here only [3], as an example, where "ultimate acceleration" is attempted by making at most five Newton steps to solve the subproblem. In fact, ideally, one would hope that one Newton step might be enough, which would then give "pure" superlinear convergence in the classical Newtonian sense. However, we are not aware of any analysis showing that one or any other fixed number of Newton steps do the job under any reasonable assumptions, and we have reasons to believe that one Newton step is not enough. Thus, our two-step development provided below.

Some words about our notation. We use $\langle \cdot, \cdot \rangle$ to denote the Euclidean inner product, $\|\cdot\|$ the associated norm, B the closed unit ball, and I the identity matrix (the space and dimensions are always clear from the context). By $|\mathcal{I}|$ we denote the cardinality of an index set $\mathcal{I}$. For any matrix M, $M_{\mathcal{I}}$ denotes the submatrix of M with rows indexed by the set $\mathcal{I}$. When in matrix notation, vectors are considered columns, and for a vector x we denote by $x_{\mathcal{I}}$ the subvector of x with coordinates indexed by $\mathcal{I}$. For a set $S \subset \mathbb{R}^q$ and a point $z \in \mathbb{R}^q$, the distance from z to S is defined by $\operatorname{dist}(z, S) = \inf_{s \in S} \|z - s\|$. Then $\Pi_S(z) = \{s \in S \mid \operatorname{dist}(z, S) = \|z - s\|\}$ is the set of all points in S that have minimal distance to z. Recall that for any closed convex set S the mapping Π_S is monotone, i.e., $\langle \Pi_S(z) - \Pi_S(w), z - w \rangle \geq 0$ for all $z, w \in \mathbb{R}^q$. For a cone $K \subset \mathbb{R}^q$, its polar (negative dual) is $K^\circ = \{\xi \in \mathbb{R}^q \mid \langle z, \xi \rangle \leq 0 \ \forall z \in K\}$. For any function $\psi : \mathbb{R}_+ \to \mathbb{R}^q$ we use the notation $\psi(t) = o(t)$ if $\lim_{t \to 0+} t^{-1}\psi(t) = 0$ and $\psi(t) = O(t)$ if $\limsup_{t \to 0+} \|t^{-1}\psi(t)\| < +\infty$.

We finish this section by recalling the superlinear convergence result of [9] for the augmented Lagrangian method (7)–(8). Our main result would be obtained by showing that after two Newton steps for the subproblem, the obtained point satisfies the needed assumptions. The following is essentially [9, Theorem 3.4 (iii)] (with the penalty parameter $\rho_k = 1/\sigma_k$), stated here in a slightly different form, more convenient for our application in the sequel.

Theorem 1 *Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be any function such that $\psi(t) = o(t)$, $\hat{c} > 0$ be a constant, and let $\{(x^k, \mu^k)\}$ be any sequence satisfying*

$$\left\| \frac{\partial \bar{L}}{\partial x}(x^{k+1}, \mu^k; \sigma_k) \right\| \leq \psi(\sigma_k), \qquad \left\| \begin{bmatrix} x^{k+1} - x^k \\ \mu^{k+1} - \mu^k \end{bmatrix} \right\| \leq \hat{c}\, \sigma_k, \tag{9}$$

where $\sigma_k = r(x^k, \mu^k)$ with r given by (3), $\mu_i^{k+1} = \mu_i^k + \frac{1}{\sigma_k} g_i(x^{k+1})$ for $i = 1, \ldots, l$ and $\mu_i^{k+1} = \max\{0, \mu_i^k + \frac{1}{\sigma_k} g_i(x^{k+1})\}$ for $i = l+1, \ldots, m$.

If $(\bar{x}, \bar{\mu})$ *satisfies SOSC* (4)*, then there exists* $\bar{\varepsilon} > 0$ *such that whenever* $(x^0, \mu^0) \in (\bar{x}, \bar{\mu}) + \bar{\varepsilon} B$ *with* $\mu_i^0 \geq 0$ *for* $i = l+1, \ldots, m$*, then the sequence* $\{(x^k, \mu^k)\}$ *converges Q-superlinearly to* $(\bar{x}, \hat{\mu})$ *with some* $\hat{\mu} \in \mathcal{M}(\bar{x})$.

We note that in the case of the equality-constrained problem (no inequality constraints), SOSC (4) can be replaced by the weaker assumption that the multiplier $\bar{\mu}$ is noncritical [17].

2 Two Newton steps for the augmented Lagrangian subproblem

We are now ready to describe our two Newton steps scheme for the augmented Lagrangian iteration. The key ingredients are the following.

The first step identifies active (inequality) constraints, including strongly and weakly active, using the technique proposed in [6], also used in [5], among many others. Note that this step is not needed in the equality-constrained case, when there are no inequalities in (1). Let $\alpha \in (0, 1)$ be fixed, and define the following index sets:

$$\mathcal{A}_1(x, \mu) = \{1, \ldots, l\} \cup \{i \in \{l+1, \ldots, m\} \mid g_i(x) \geq -(r(x, \mu))^\alpha,\ \mu_i \geq (r(x, \mu))^\alpha\},$$
$$\mathcal{A}_0(x, \mu) = \{i \in \{l+1, \ldots, m\} \mid g_i(x) \geq -(r(x, \mu))^\alpha,\ \mu_i < (r(x, \mu))^\alpha\},$$

where r is the natural residual of the KKT system (2), defined in (3). Under the error bound condition (6), locally the second part in $\mathcal{A}_1(x, \mu)$ correctly identifies strongly active constraints (inequality constraints active at $\bar{x}$ with the associated multiplier $\bar{\mu}_i$ positive), while $\mathcal{A}_0(x, \mu)$ identifies weakly active constraints (those with zero multiplier), see [6]. In particular, this holds true under SOSC (4), since it implies the error bound.

Having $\mathcal{A}_1(x, \mu)$ and $\mathcal{A}_0(x, \mu)$, a QP is solved whose structure is closely related to that of the subproblem of stabilized sequential quadratic programming (sSQP) [8,10,13,23], [18, Chapter 7.2]. The only difference is that inequality constraints in the set $\mathcal{A}_1(x, \mu)$ are treated as equalities. After a stationary point of this QP is computed, the multiplier estimates are updated as in the usual augmented Lagrangian algorithm.

Next, a second QP is solved. It again has the same sSQP structure, but now the constraints are linearized at the primal solution of the first QP, similarly to what is called second-order corrections in SQP literature; see, e.g., [18, Chapter 4.3.6]. Another feature to point out is that the Hessian of the Lagrangian in the second QP is computed at the old iterate x^k but with the new multiplier estimates obtained after solving the first QP. The primal part of a stationary point of the second QP gives the next iterate x^{k+1}, with μ^{k+1} obtained by the usual augmented Lagrangian update using the old iterate μ^k (not the intermediate estimates obtained after the first QP).

Specifically, the procedure is the following.

Algorithm 1 Choose $\alpha \in (0, 1)$ and $(x^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^m$ with $\mu_i^0 \geq 0$ for $i = l+1, \ldots, m$. Set $k = 0$.

1. If $r(x^k, \mu^k) = 0$, STOP.
   Else, define $\mathcal{A}_1 = \mathcal{A}_1(x^k, \mu^k)$, $\mathcal{A}_0 = \mathcal{A}_0(x^k, \mu^k)$, $\sigma_k = r(x^k, \mu^k)$.

2. Find a stationary point $(\bar{y}, \bar{\nu})$ of

$$\begin{aligned}\underset{(y,\nu)}{\text{minimize}}\ & \langle f'(x^k), y - x^k\rangle + \frac{1}{2}\left\langle \frac{\partial^2 L}{\partial x^2}(x^k, \mu^k)(y - x^k), y - x^k\right\rangle + \frac{\sigma_k}{2}\|\nu\|^2\\ \text{s.t. } & g_i(x^k) + \langle g_i'(x^k), y - x^k\rangle - \sigma_k(\nu_i - \mu_i^k) = 0,\ i \in \mathcal{A}_1,\\ & g_i(x^k) + \langle g_i'(x^k), y - x^k\rangle - \sigma_k(\nu_i - \mu_i^k) \le 0,\ i \in \mathcal{A}_0.\end{aligned}$$

Define $\tilde{x}^k = \bar{y}$, $\tilde{\mu}_i^k = \mu_i^k + \frac{1}{\sigma_k} g_i(\tilde{x}^k)$ for $i = 1, \ldots, l$,
and $\tilde{\mu}_i^k = \max\{0, \mu_i^k + \frac{1}{\sigma_k} g_i(\tilde{x}^k)\}$ for $i = l + 1, \ldots, m$.

3. Find a stationary point $(\bar{y}, \bar{\nu})$ of

$$\begin{aligned}\underset{(y,\nu)}{\text{minimize}}\ & \langle f'(x^k), y - x^k\rangle + \frac{1}{2}\left\langle \frac{\partial^2 L}{\partial x^2}(x^k, \tilde{\mu}^k)(y - x^k), y - x^k\right\rangle + \frac{\sigma_k}{2}\|\nu\|^2\\ \text{s.t. } & g_i(\tilde{x}^k) + \langle g_i'(x^k), y - \tilde{x}^k\rangle - \sigma_k(\nu_i - \mu_i^k) = 0,\ i \in \mathcal{A}_1,\\ & g_i(\tilde{x}^k) + \langle g_i'(x^k), y - \tilde{x}^k\rangle - \sigma_k(\nu_i - \mu_i^k) \le 0,\ i \in \mathcal{A}_0.\end{aligned}$$

Define $x^{k+1} = \bar{y}$, $\mu_i^{k+1} = \mu_i^k + \frac{1}{\sigma_k} g_i(x^{k+1})$ for $i = 1, \ldots, l$,
and $\mu_i^{k+1} = \max\{0, \mu_i^k + \frac{1}{\sigma_k} g_i(x^{k+1})\}$ for $i = l + 1, \ldots, m$.

4. Set $k = k + 1$ and go to step 1.

Note that in the equality-constrained case [no inequality constraints in (1)], the two QPs in Algorithm 1 are also equality-constrained. In that case, computing a stationary point and an associated Lagrange multiplier of a QP is equavalent to solving a system of linear equations. Thus, in the equality-constrained case, computing a suitable approximate solution of the augmented Lagrangian subproblem would be achieved by solving just two linear equations. Note also that in this case, the task of indices identification in Step 1 is not needed.

As Algorithm 1 is related (in part) to sSQP, we next give a few brief remarks on the latter. First, sSQP is a "one-step" Newtonian method, locally superlinearly convergent under SOSC only [8] (in the equality-constrained case, even under the weaker assumption of noncriticality of the Lagrange multiplier, see [18, Chapter 7.2]). However, while globally convergent algorithms which attempt to use sSQP subproblems do exist [7,11,12,19,20], those constructions are rather complicated and have certain drawbacks. That said, in [11,12] it appears that solving one sSQP subproblem is locally enough within a globally convergent *primal–dual augmented Lagrangian* algorithm (primal–dual augmented Lagrangian function is different from the augmented Lagrangian). In the case of the usual augmented Lagrangian method, we need two QP subproblems (Algorithm 1).

3 Convergence analysis

We shall show that the iterates of Algorithm 1 satisfy the conditions of Theorem 1. This would imply that solving two QPs is enough to produce an appropriate approximate solution of the augmented Lagrangian subproblem.

The first order of business is to prove that the procedure is well-defined, i.e., that the two QPs have solutions.

To simplify the notation, for each (x, μ) we shall use $\mathcal{A}$ to denote the index set given by

$$\mathcal{A}(x, \mu) = \mathcal{A}_0(x, \mu) \cup \mathcal{A}_1(x, \mu).$$

In the sequel, the points where $\mathcal{A}$ is evaluated are always clear from the context. The identification of active constraints follows from [6, Theorem 2.2]. For completeness, we state this result (in a form adapted for our purposes).

Theorem 2 *Let $\bar{x}$ be an isolated stationary point of problem (1), $\bar{\mu} \in \mathcal{M}(\bar{x})$, and let $\rho : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_+$ be a continuous function, zero-valued on $\{\bar{x}\} \times \mathcal{M}(\bar{x})$, and satisfying*

$$\lim_{\substack{(x,\mu)\to(\bar{x},\bar{\mu}) \\ (x,\mu)\notin\{\bar{x}\}\times\mathcal{M}(\bar{x})}} \frac{\rho(x, \mu)}{\|x - \bar{x}\| + dist(\mu, \mathcal{M}(\bar{x}))} = +\infty.$$

Then, there exists a neighborhood $\mathcal{V}$ of $(\bar{x}, \bar{\mu})$ such that

$$\{i \in \{l+1, \dots, m\} \mid g_i(x) \geq -\rho(x, \mu)\} = \{i \in \{l+1, \dots, m\} \mid g_i(\bar{x}) = 0\},$$

for any $(x, \mu) \in \mathcal{V}$.

If $(\bar{x}, \bar{\mu})$ satisfies SOSC (4), then the primal part $\bar{x}$ of the solution is locally unique and the error bound (6) holds. Thus, by Theorem 2, taking $\rho(x, \mu) = (r(x, \mu))^\alpha$, $\alpha \in (0, 1)$, we obtain that for any (x, μ) close enough to $(\bar{x}, \bar{\mu})$ it holds that $\mathcal{A}(x, \mu) = \mathcal{I}$, where

$$\mathcal{I} = \mathcal{I}(\bar{x}) = \{i \in \{1, \dots, m\} \mid g_i(\bar{x}) = 0\}$$

is the index set of all equality constraints and of inequality constraints active at $\bar{x}$.

The following auxiliary result is essential to proving solvability of the two QPs in Algorithm 1; it is related to [8, Proposition 1].

Proposition 1 *Assume that SOSC (4) holds at $(\bar{x}, \bar{\mu})$. Then there exist constants $\gamma_c > 0$ and $\varepsilon_c > 0$ such that if $(y, \nu) \in (\bar{x}, \bar{\mu}) + \varepsilon_c B$ and $(x, \mu) \in (\bar{x}, \bar{\mu}) + \varepsilon_c B$, it holds that*

$$\left\langle \frac{\partial^2 L}{\partial x^2}(y, \nu)u, u \right\rangle + r(x, \mu)\|v\|^2 \geq \gamma_c \Big(\|u\|^2 + r(x, \mu)\|v\|^2 \Big) \tag{10}$$

for all $(u, v) \in \hat{\mathcal{C}}(x, \mu)$, where r is defined in (3) and

$$\hat{\mathcal{C}}(x, \mu) = \left\{ (u, v) \in \mathbb{R}^n \times \mathbb{R}^{|\mathcal{A}|} \;\middle|\; \begin{array}{l} \langle g_i'(x), u \rangle = r(x, \mu)v_i, \ i \in \mathcal{A}_1(x, \mu) \\ \langle g_i'(x), u \rangle \leq r(x, \mu)v_i, \ i \in \mathcal{A}_0(x, \mu) \end{array} \right\}. \tag{11}$$

Proof Suppose the contrary, i.e., that there exist $\{(y^k, \nu^k)\}$, $\{(x^k, \mu^k)\}$ converging to $(\bar{x}, \bar{\mu})$ and $(u^k, v^k) \in \hat{\mathcal{C}}(x^k, \mu^k)$ such that

$$\left\langle \frac{\partial^2 L}{\partial x^2}(y^k, \nu^k)u^k, u^k \right\rangle + r_k\|v^k\|^2 < \frac{1}{k}(\|u^k\|^2 + r_k\|v^k\|^2), \tag{12}$$

where $r_k = r(x^k, \mu^k)$. In particular, $r_k \to 0$. Evidently, (12) subsumes that $(u^k, v^k) \neq 0$. Define $\eta_k = \|(u^k, \sqrt{r_k}v^k)\| > 0$. Passing onto a subsequence, if necessary, we can assume that

$$\frac{1}{\eta_k}\begin{bmatrix} u^k \\ \sqrt{r_k}v^k \end{bmatrix} \to \begin{bmatrix} \bar{u} \\ \bar{w} \end{bmatrix} \neq 0. \tag{13}$$

By [6, Theorem 2.2], $\mathcal{A}(x^k, \mu^k) = \mathcal{I}$ for all k large enough. Since $\mathcal{A}_1(x^k, \mu^k) \cup \mathcal{A}_0(x^k, \mu^k) = \mathcal{A}(x^k, \mu^k) = \mathcal{I}$, passing onto a subsequence if necessary, we can assume that $\mathcal{A}_1(x^k, \mu^k)$ and $\mathcal{A}_0(x^k, \mu^k)$ are fixed index sets, which we shall denote by $\mathcal{A}_1$ and $\mathcal{A}_0$, respectively.

Observe that since $r_k \to 0$ and $\sqrt{r_k}v^k/\eta_k$ is bounded, it holds that

$$r_k\frac{v^k}{\eta_k} = \sqrt{r_k}\frac{\sqrt{r_k}v^k}{\eta_k} \to 0. \tag{14}$$

Since $\hat{\mathcal{C}}(x^k, \mu^k)$ is a cone and $(u^k, v^k) \in \hat{\mathcal{C}}(x^k, \mu^k)$, we conclude that $(u^k/\eta_k, v^k/\eta_k) \in \hat{\mathcal{C}}(x^k, \mu^k)$. Dividing now relations in (11) by η_k, passing onto the limit and taking into account (14), we obtain that

$$\langle g_i'(\bar{x}), \bar{u}\rangle = 0 \ \forall\, i \in \mathcal{A}_1, \quad \langle g_i'(\bar{x}), \bar{u}\rangle \leq 0 \ \forall\, i \in \mathcal{A}_0. \tag{15}$$

By the definition of $\mathcal{A}_1$ we conclude that

$$\{1, \ldots, l\} \cup \{i \in \{l+1, \ldots, m\} \mid \bar{\mu}_i > 0\} \subset \mathcal{A}_1. \tag{16}$$

Then, from the KKT relations (2), using also (15) and (16), we obtain that

$$0 = \langle f'(\bar{x}) + (g'(\bar{x}))^\top \bar{\mu}, \bar{u}\rangle = \langle f'(\bar{x}), \bar{u}\rangle + \sum_{i=1}^{m} \bar{\mu}_i \langle g_i'(\bar{x}), \bar{u}\rangle = \langle f'(\bar{x}), \bar{u}\rangle.$$

Together with (15), this shows that $\bar{u} \in \mathcal{C}$.

On the other hand, dividing (12) by η_k^2 and taking limits, we obtain that

$$\left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\mu})\bar{u}, \bar{u} \right\rangle + \|\bar{w}\|^2 \leq 0. \tag{17}$$

This shows that $\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\mu})\bar{u}, \bar{u}\rangle \leq 0$ for $\bar{u} \in \mathcal{C}$. By SOSC (4), it then follows that $\bar{u} = 0$. But then, from (17), we also have $\bar{w} = 0$. This contradicts (13). □

For a fixed (x, μ), define the cone

$$Q_{\mathcal{A}} = \left\{ \nu \in \mathbb{R}^{|\mathcal{A}|} \mid \nu_i \in \mathbb{R}, i \in \mathcal{A}_1(x, \mu);\ \nu_i \geq 0, i \in \mathcal{A}_0(x, \mu) \right\}.$$

Note that by the structure of $Q_{\mathcal{A}}$, it holds that

$$\nu = \Pi_{Q_{\mathcal{A}}}(\lambda) \ \Leftrightarrow\ \nu_i = \lambda_i \text{ for } i \in \mathcal{A}_1(x, \mu),\ \ \nu_i = \max\{0, \lambda_i\} \text{ for } i \in \mathcal{A}_0(x, \mu). \tag{18}$$

For parameters $(x, \mu) \in \mathbb{R}^n \times \mathbb{R}^m$, $(\hat{x}, \hat{\mu}) \in \mathbb{R}^n \times \mathbb{R}^m$, consider the function $\phi : \mathbb{R}^n \to \mathbb{R}$ given by

$$\begin{aligned}\phi(y; x, \mu, \hat{x}, \hat{\mu}) = \langle f'(x), y - x\rangle &+ \frac{1}{2}\left\langle \frac{\partial^2 L}{\partial x^2}(x, \hat{\mu})(y - x), y - x \right\rangle \\ &+ \frac{r}{2}\left\| \Pi_{Q_{\mathcal{A}}}\left(\mu_{\mathcal{A}} + \frac{1}{r} g_{\mathcal{A}}(\hat{x}) + \frac{1}{r} g'_{\mathcal{A}}(x)(y - \hat{x}) \right) \right\|^2,\end{aligned}$$

where $r = r(x, \mu)$ defined in (3). The function ϕ is piecewise quadratic and continuously differentiable. Moreover, stationary points for the QPs in Algorithm 1 can be obtained from stationary points of ϕ.

Proposition 2 *The quadratic problem*

$$\begin{aligned}\underset{(y,\nu)}{\text{minimize}}\ \ &\langle f'(x), y - x\rangle + \frac{1}{2}\left\langle \frac{\partial^2 L}{\partial x^2}(x, \hat{\mu})(y - x), y - x \right\rangle + \frac{r}{2}\|\nu\|^2 \\ \text{s.t.}\ \ &g_i(\hat{x}) + \langle g'_i(x), y - \hat{x}\rangle - r(\nu_i - \mu_i) = 0,\ i \in \mathcal{A}_1, \\ &g_i(\hat{x}) + \langle g'_i(x), y - \hat{x}\rangle - r(\nu_i - \mu_i) \leq 0,\ i \in \mathcal{A}_0,\end{aligned}$$

has a stationary point $(\bar{y}, \bar{\nu})$ if and only if

$$0 = \phi'(\bar{y}; x, \mu, \hat{x}, \hat{\mu}) \quad \textit{and} \quad \bar{\nu} = \Pi_{Q_{\mathcal{A}}}\left(\mu_{\mathcal{A}} + \frac{1}{r} g_{\mathcal{A}}(\hat{x}) + \frac{1}{r} g'_{\mathcal{A}}(x)(\bar{y} - \hat{x}) \right).$$

Proof First, note that

$$\begin{aligned}\phi'(y; x, \mu, \hat{x}, \hat{\mu}) = &f'(x) + (g'_{\mathcal{A}}(x))^\top \Pi_{Q_{\mathcal{A}}}\left(\mu_{\mathcal{A}} + \tfrac{1}{r} g_{\mathcal{A}}(\hat{x}) + \tfrac{1}{r} g'_{\mathcal{A}}(x)(y - \hat{x}) \right) \\ &+ \frac{\partial^2 L}{\partial x^2}(x, \hat{\mu})(y - x).\end{aligned}$$

Thus, the two equalities

$$0 = \phi'(\bar{y}; x, \mu, \hat{x}, \hat{\mu}),$$
$$\bar{\nu} = \Pi_{Q_{\mathcal{A}}} \left(\mu_{\mathcal{A}} + \frac{1}{r} g_{\mathcal{A}}(\hat{x}) + \frac{1}{r} g'_{\mathcal{A}}(x)(\bar{y} - \hat{x}) \right),$$

hold if and only if

$$\begin{aligned} 0 &= f'(x) + \frac{\partial^2 L}{\partial x^2}(x, \hat{\mu})(\bar{y} - x) + (g'_{\mathcal{A}}(x))^\top \bar{\nu}, \\ \bar{\nu}_i &= \mu_i + \frac{1}{r} g_i(\hat{x}) + \frac{1}{r} \langle g'_i(x), \bar{y} - \hat{x} \rangle, \quad i \in \mathcal{A}_1, \\ \bar{\nu}_i &= \max \left\{ 0, \mu_i + \frac{1}{r} g_i(\hat{x}) + \frac{1}{r} \langle g'_i(x), \bar{y} - \hat{x} \rangle \right\}, \quad i \in \mathcal{A}_0, \end{aligned}$$

where (18) was taken into account. The last two equations can be written as

$$\begin{aligned} & g_i(\hat{x}) + \langle g'_i(x), \bar{y} - \hat{x} \rangle - r(\bar{\nu}_i - \mu_i) = 0, \; i \in \mathcal{A}_1, \\ & \left. \begin{array}{l} 0 \le \bar{\nu}_i, \; g_i(\hat{x}) + \langle g'_i(x), \bar{y} - \hat{x} \rangle - r(\bar{\nu}_i - \mu_i) \le 0, \\ \bar{\nu}_i \left(g_i(\hat{x}) + \langle g'_i(x), \bar{y} - \hat{x} \rangle - r(\bar{\nu}_i - \mu_i) \right) = 0, \end{array} \right. \; i \in \mathcal{A}_0. \end{aligned}$$

It is then easy to see that the relations above are equivalent to the KKT conditions of the quadratic problem in question, which are:

$$\begin{aligned} & 0 = f'(x) + \frac{\partial^2 L}{\partial x^2}(x, \hat{\mu})(\bar{y} - x) + (g'_{\mathcal{A}}(x))^\top \lambda, \\ & 0 = r\bar{\nu} - r\lambda, \\ & 0 = g_i(\hat{x}) + \langle g'_i(x), \bar{y} - \hat{x} \rangle - r(\bar{\nu}_i - \mu_i), \; i \in \mathcal{A}_1, \\ & \left. \begin{array}{l} 0 \le \lambda_i, \; g_i(\hat{x}) + \langle g'_i(x), \bar{y} - \hat{x} \rangle - r(\bar{\nu}_i - \mu_i) \le 0, \\ \lambda_i \left(g_i(\hat{x}) + \langle g'_i(x), \bar{y} - \hat{x} \rangle - r(\bar{\nu}_i - \mu_i) \right) = 0, \end{array} \right. \; i \in \mathcal{A}_0. \end{aligned}$$

□

Hence, for the QPs in Algorithm 1, $(\bar{y}, \bar{\nu})$ is a stationary point of the QP in step 2 if and only if $0 = \phi'(\bar{y}; x^k, \mu^k, x^k, \mu^k)$, and $(\bar{y}, \bar{\nu})$ is a stationary point of the QP in step 3 if and only if $0 = \phi'(\bar{y}; x^k, \mu^k, \tilde{x}^k, \tilde{\mu}^k)$. To show the solvability of the two QPs in Algorithm 1, we shall prove the existence of stationary points of ϕ, showing the existence of its minimizers.

Proposition 3 *For any $x \in \mathbb{R}^n$, $\hat{x} \in \mathbb{R}^n$ and $\mu \in \mathbb{R}^m$, $\hat{\mu} \in \mathbb{R}^m$ such that $(x, \mu) \in (\bar{x}, \bar{\mu}) + \varepsilon_c B$ and $(x, \hat{\mu}) \in (\bar{x}, \bar{\mu}) + \varepsilon_c B$, the function $\phi(\cdot\,; x, \mu, \hat{x}, \hat{\mu})$ is coercive.*

Hence, this function has a minimizer, and thus a stationary point. Consequently, the two QPs in Algorithm 1 have stationary points, and Algorithm 1 is well-defined.

Proof Suppose that ϕ is not coercive, i.e., there exist a sequence $\{y^k\} \subset \mathbb{R}^n$ with $\|y^k\| \to +\infty$ and $\beta \in \mathbb{R}$ such that $\phi(y^k; x, \mu; \hat{x}, \hat{\mu}) \le \beta$ for all k.

Taking a subsequence, if necessary, we can assume that

$$\frac{y^k}{\|y^k\|} \to u \neq 0. \tag{19}$$

Then,

$$\begin{aligned}0 &= \lim_{k\to+\infty} \frac{\beta}{\|y^k\|^2}\\ &\geq \lim_{k\to+\infty} \frac{1}{\|y^k\|^2}\phi(y^k; x, \mu, \hat{x}, \hat{\mu})\\ &= \frac{1}{2}\left\langle \frac{\partial^2 L}{\partial x^2}(x, \hat{\mu})u, u\right\rangle + \frac{r}{2}\left\| \Pi_{Q_{\mathcal{A}}}\left(\frac{1}{r}g'_{\mathcal{A}}(x)u\right)\right\|^2.\end{aligned}$$

Define $v = \Pi_{Q_{\mathcal{A}}}\left(\frac{1}{r}g'_{\mathcal{A}}(x)u\right)$. Then, taking into account (18), it is seen that $(u, v) \in \hat{\mathcal{C}}(x, \mu)$, where $\hat{\mathcal{C}}(x, \mu)$ is defined in (11). Hence, by Proposition 1 [in particular, (10)], it holds that

$$0 \geq \left\langle \frac{\partial^2 L}{\partial x^2}(x, \hat{\mu})u, u\right\rangle + r\|v\|^2 \geq \gamma_c(\|u\|^2 + r\|v\|^2) \geq \gamma_c\|u\|^2.$$

This contradicts (19). $\square$

Next, we show that the primal–dual sequence generated by Algorithm 1 satisfies the conditions in Theorem 1 [in particular, the relations in (9)]. We start with a chain of results that end up providing a bound for the distance between succesive iterates, needed in Theorem 1.

Recall that the error bound (6) holds. In particular, if $\check{\mu}$ is the projection of μ onto $\mathcal{M}(\bar{x})$, for any $i \notin \mathcal{I}$ we have

$$|\mu_i| = |\mu_i - \check{\mu}_i| \leq \operatorname{dist}(\mu, \mathcal{M}(\bar{x})) \leq \beta\, r(x, \mu), \tag{20}$$

where we use the fact that if $\check{\mu} \in \mathcal{M}(\bar{x})$ then $\check{\mu}_i = 0$ for $i \notin \mathcal{I}$.

Proposition 4 *Let $(\bar{x}, \bar{\mu})$ satisfy SOSC (4). Let $\hat{x} = \hat{x}(x, \mu)$ be such that $\hat{x} - x = O(r(x, \mu))$, and $\hat{\mu} = \hat{\mu}(x, \mu)$ be such that $\lim_{(x,\mu)\to(\bar{x},\bar{\mu})} \hat{\mu} = \bar{\mu}$.*

Then there exist $\varepsilon_e, c_e > 0$ such that if the point $(x, \mu) \in (\bar{x}, \bar{\mu}) + \varepsilon_e B$ satisfies $r = r(x, \mu) > 0$ and $\mu_i \geq 0$ for $i = l+1, \ldots, m$, and if y satisfies $\phi'(y; x, \mu, \hat{x}, \hat{\mu}) = 0$, it holds that

$$\|y - x\| + \left\|\Pi_{Q_{\mathcal{A}}}\left(\mu_{\mathcal{A}} + \tfrac{1}{r}g_{\mathcal{A}}(\hat{x}) + \tfrac{1}{r}g'_{\mathcal{A}}(x)(y - \hat{x})\right) - \mu_{\mathcal{A}}\right\| \leq c_e\, r. \tag{21}$$

Proof Suppose the contrary to the assertion. Then there exist sequences $\{(x^k, \mu^k)\}$, $\{(\hat{x}^k, \hat{\mu}^k)\}$ and $\{y^k\}$ such that $(x^k, \mu^k) \to (\bar{x}, \bar{\mu})$, $r_k = r(x^k, \mu^k) > 0$, $\mu_i^k \geq 0$ for $i = l+1, \ldots, m$, $\{(\hat{x}^k - x^k)/r_k\}$ is bounded, $\hat{\mu}^k \to \bar{\mu}$, $0 = \phi'(y^k; x^k, \mu^k, \hat{x}^k, \hat{\mu}^k)$, and they satisfy

$$\frac{r_k}{\eta_k} \to 0, \tag{22}$$

where (using the fact that $\mathcal{A}(x^k, \mu^k) = \mathcal{I}$ for k large enough), we denote $\eta_k = \|y^k - x^k\| + \|\nu_{\mathcal{I}}^k - \mu_{\mathcal{I}}^k\|$ with

$$\nu_{\mathcal{I}}^k = \Pi_{Q_{\mathcal{A}}}\left(\mu_{\mathcal{I}}^k + \frac{1}{r_k} g_{\mathcal{I}}(\hat{x}^k) + \frac{1}{r_k} g'_{\mathcal{I}}(x^k)(y^k - \hat{x}^k)\right). \tag{23}$$

Taking a subsequence, if necessary, we can assume that

$$\frac{1}{\eta_k}\begin{bmatrix} y^k - x^k \\ \nu_{\mathcal{I}}^k - \mu_{\mathcal{I}}^k \end{bmatrix} \to \begin{bmatrix} u \\ v_{\mathcal{I}} \end{bmatrix} \neq 0. \tag{24}$$

Also, since $\mathcal{A}_0(x^k, \mu^k) \cup \mathcal{A}_1(x^k, \mu^k) = \mathcal{I}$ is finite, we can assume that $\mathcal{A}_1(x^k, \mu^k)$ and $\mathcal{A}_0(x^k, \mu^k)$ are some fixed index sets, say $\mathcal{A}_1$ and $\mathcal{A}_0$.

We first show that $u = 0$. By (23) and (18), we have that

$$\begin{aligned} g_i(\hat{x}^k) + g'_i(x^k)(y^k - \hat{x}^k) - r_k(\nu_i^k - \mu_i^k) &= 0, \quad \forall i \in \mathcal{A}_1, \\ g_i(\hat{x}^k) + g'_i(x^k)(y^k - \hat{x}^k) - r_k(\nu_i^k - \mu_i^k) &\le 0, \quad \forall i \in \mathcal{A}_0. \end{aligned} \tag{25}$$

Since $\hat{x}^k - x^k = O(r_k)$ by the assumption, $x^k - \bar{x} = O(r_k)$ by the error bound (6), and $r_k = o(\eta_k)$ by (22), we conclude that

$$y^k - \hat{x}^k = y^k - x^k + x^k - \hat{x}^k = y^k - x^k + o(\eta_k),$$

and for $i \in \mathcal{I}$,

$$g_i(\hat{x}^k) = g_i(\hat{x}^k) - g_i(x^k) + g_i(x^k) - g_i(\bar{x}) = o(\eta_k).$$

Then, dividing by η_k in (25) and taking limits, we obtain that

$$\langle g'_i(\bar{x}), u\rangle = 0 \quad \forall i \in \mathcal{A}_1, \qquad \langle g'_i(\bar{x}), u\rangle \le 0 \quad \forall i \in \mathcal{A}_0.$$

Then, the same way as in the proof of Propistion 1 [after (15) therein was obtained], we conclude that $u \in \mathcal{C}$.

By the definition of r in (3), it holds that

$$f'(x^k) + (g'(x^k))^\top \mu^k = O(r_k) = o(\eta_k),$$

and by (20) we have $|\mu_i^k| = O(r_k) = o(\eta_k)$ for any $i \notin \mathcal{I}$. Hence,

$$\begin{aligned} 0 = \phi'(y^k; x^k, \mu^k, \hat{x}^k, \hat{\mu}^k) &= f'(x^k) + \frac{\partial^2 L}{\partial x^2}(x^k, \hat{\mu}^k)(y^k - x^k) + (g'_{\mathcal{I}}(x^k))^\top \nu_{\mathcal{I}}^k \\ &= \frac{\partial^2 L}{\partial x^2}(x^k, \hat{\mu}^k)(y^k - x^k) + (g'_{\mathcal{I}}(x^k))^\top(\nu_{\mathcal{I}}^k - \mu_{\mathcal{I}}^k) + o(\eta_k). \end{aligned} \tag{26}$$

Since $\mu_{\mathcal{I}}^k \in Q_{\mathcal{A}}$, by (23) and by the monotonicity of the projection mapping, it holds that

$$\langle \nu_{\mathcal{I}}^k - \mu_{\mathcal{I}}^k, g_{\mathcal{I}}(\hat{x}^k) + g'_{\mathcal{I}}(x^k)(y^k - \hat{x}^k)\rangle \geq 0.$$

Using that $y^k - \hat{x}^k = O(\eta_k)$, $\nu_{\mathcal{I}}^k - \mu_{\mathcal{I}}^k = O(\eta_k)$ and $g_{\mathcal{I}}(\hat{x}^k) = o(\eta_k)$, we obtain from (26) that

$$\begin{aligned} 0 &= \langle \phi'(y^k; x^k, \mu^k, \hat{x}^k, \hat{\mu}^k), y^k - \hat{x}^k\rangle \\ &\geq \left\langle \frac{\partial^2 L}{\partial x^2}(x^k, \hat{\mu}^k)(y^k - x^k), y^k - \hat{x}^k \right\rangle + o(\eta_k^2) \\ &= \left\langle \frac{\partial^2 L}{\partial x^2}(x^k, \hat{\mu}^k)(y^k - x^k), y^k - x^k \right\rangle + o(\eta_k^2), \end{aligned}$$

where the facts that $y^k - \hat{x}^k = y^k - x^k + o(\eta_k)$ and $y^k - x^k = O(\eta_k)$ were taken into account for the last equality. Dividing the inequality above by η_k^2, taking limits and using that $\hat{\mu}^k \to \bar{\mu}$, yields

$$0 \geq \left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\mu})u, u \right\rangle.$$

Since $u \in \mathcal{C}$, SOSC (4) implies that $u = 0$.

Now, dividing (26) by η_k, taking limits and using that $u = 0$ yields

$$v_{\mathcal{I}} \in K = \ker\, (g'_{\mathcal{I}}(\bar{x}))^{\top}.$$

Using that, by the assumption and by the error bound (6),

$$\hat{x}^k - \bar{x} = \hat{x}^k - x^k + x^k - \bar{x} = O(r_k),$$

and that $r_k = o(\eta_k)$, we obtain

$$g_{\mathcal{I}}(\hat{x}^k) = g_{\mathcal{I}}(\bar{x}) + g'_{\mathcal{I}}(\bar{x})(\hat{x}^k - \bar{x}) + O(\|\hat{x}^k - \bar{x}\|^2) = g'_{\mathcal{I}}(\bar{x})(\hat{x}^k - \bar{x}) + o(r_k\eta_k),$$

and

$$g'_{\mathcal{I}}(x^k)(y^k - \hat{x}^k) = g'_{\mathcal{I}}(\bar{x})(y^k - \hat{x}^k) + o(r_k\eta_k).$$

Adding the last two relations, it follows that

$$g_{\mathcal{I}}(\hat{x}^k) + g'_{\mathcal{I}}(x^k)(y^k - \hat{x}^k) + o(r_k\eta_k) = g'_{\mathcal{I}}(\bar{x})(y^k - \bar{x}) \in \operatorname{im} g'_{\mathcal{I}}(\bar{x}) = K^{\perp}.$$

Then, using the linearity of Π_K (since K is a subpspace) and closedness of $K^{\perp}$, it follows that

$$\frac{1}{r_k\eta_k}\Pi_K\left(g_{\mathcal{I}}(\hat{x}^k) + g'_{\mathcal{I}}(x^k)(y^k - \hat{x}^k)\right) \to 0. \tag{27}$$

Consider the indices $i \in \mathcal{A}_0$ with $\mu_i^k + \frac{1}{r_k} g_i(\hat{x}^k) + \frac{1}{r_k}\langle g_i'(x^k), y^k - \hat{x}^k\rangle < 0$. Passing onto a subsequence if necessary, we can assume that there exists a fixed set $\mathcal{A}_\ell \subseteq \mathcal{A}_0$ such that the latter inequality holds for all $i \in \mathcal{A}_\ell$ and all k (note that $\mathcal{A}_\ell$ can be empty). Then, by (18),

$$\nu_i^k = 0 \text{ for } i \in \mathcal{A}_\ell \text{ and } \nu_i^k = \mu_i^k + \frac{1}{r_k} g_i(\hat{x}^k) + \frac{1}{r_k}\langle g_i'(x^k), y^k - \hat{x}^k\rangle \text{ for } i \notin \mathcal{A}_\ell.$$

Define the cone

$$Q_\ell = \{w_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|} \mid w_i \in \mathbb{R}, i \in \mathcal{I}\backslash\mathcal{A}_\ell;\ w_i \geq 0, i \in \mathcal{A}_\ell\}.$$

Then,

$$-(\nu_{\mathcal{I}}^k - \mu_{\mathcal{I}}^k) \in Q_\ell \quad \text{and} \quad \frac{1}{r_k} g_{\mathcal{I}}(\hat{x}^k) + \frac{1}{r_k} g_{\mathcal{I}}'(x^k)(y^k - \hat{x}^k) - (\nu_{\mathcal{I}}^k - \mu_{\mathcal{I}}^k) \in Q_\ell^\circ.$$

Dividing those relations by η_k and taking limits, we obtain that

$$-v_{\mathcal{I}} \in Q_\ell \quad \text{and} \quad -\Pi_K(v_{\mathcal{I}}) \in \{z \mid z = \Pi_K(\xi_{\mathcal{I}}) \text{ for some } \xi_{\mathcal{I}} \in Q_\ell^\circ\},$$

where we use closedness of Q_ℓ and of its polar, (27), and the linearity of $\Pi_K(\cdot)$. Since $v_{\mathcal{I}} \in K$, it follows that there exists $\xi_{\mathcal{I}} \in Q_\ell^\circ$ such that $-v_{\mathcal{I}} = \Pi_K(\xi_{\mathcal{I}})$. Then,

$$0 \geq \langle -v_{\mathcal{I}}, \xi_{\mathcal{I}}\rangle = \langle -v_{\mathcal{I}}, \Pi_K(\xi_{\mathcal{I}}) + \Pi_{K^\perp}(\xi_{\mathcal{I}})\rangle = \langle -v_{\mathcal{I}}, \Pi_K(\xi_{\mathcal{I}})\rangle = \|v_{\mathcal{I}}\|^2,$$

where we use that $-v_{\mathcal{I}} \in Q_\ell$, $\xi_{\mathcal{I}} \in Q_\ell^\circ$ and $v_{\mathcal{I}} \in K$, $\Pi_K(\xi_{\mathcal{I}}) = -v_{\mathcal{I}}$. Combining this relation with $u = 0$, we obtain that $(u, v_{\mathcal{I}}) = 0$, in contradiction with (24). ◻

The next result gives a suitable bound on the primal–dual step. We shall use [14, Lemma 2] (see also [18, Proposition 1.4.3]), which can be stated as follows.

Theorem 3 *Let $(\bar{x}, \bar{\mu})$ satisfy SOSC (4). Then, there exist $\tau > 0$ and a neighborhood $\mathcal{V}$ of $(\bar{x}, \bar{\mu})$ such that for any $p = (q, s, t) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^{m-l}$ and any $(x(p), \mu(p)) \in \mathcal{V}$, where $(x(p), \mu(p))$ is a solution of the perturbed the KKT system (2):*

$$\begin{aligned}
&0 = \frac{\partial L}{\partial x}(x, \mu) + q,\\
&0 = g_i(x) + s_i, \quad i = 1, \ldots, l\\
&0 \leq \mu_i,\ g_i(x) + t_i \leq 0,\ \mu_i(g_i(x) + t_i) = 0, \quad i = l+1, \ldots, m,
\end{aligned} \tag{28}$$

it holds that

$$\|x(p) - \bar{x}\| + dist(\mu(p), \mathcal{M}(\bar{x})) \leq \tau\|p\|. \tag{29}$$

We then obtain the following.

Proposition 5 *Under the assumptions of Proposition 4, for (x, μ) close enough to $(\bar{x}, \bar{\mu})$ and such that $r = r(x, \mu) > 0$, $\mu_i \geq 0$ for $i = l+1, \ldots, m$, and (y, λ) satisfying $\phi'(y; x, \mu, \hat{x}, \hat{\mu}) = 0$, $\lambda_i = \mu_i + \frac{1}{r} g_i(y)$ for $i = 1, \ldots, l$ and $\lambda_i = \max\{0, \mu_i + \frac{1}{r} g_i(y)\}$ for $i = l+1, \ldots, m$, it holds that*

$$\|y - x\| + \|\lambda - \mu\| = O(r(x, \mu)). \tag{30}$$

Moreover, if $\nu_i = \mu_i + \frac{1}{r} g_i(\hat{x}) + \frac{1}{r} \langle g_i'(x), y - \hat{x} \rangle$ for $i = 1, \ldots, l$ and $\nu_i = \max\{0, \mu_i + \frac{1}{r} g_i(\hat{x}) + \frac{1}{r} \langle g_i'(x), y - \hat{x} \rangle\}$ for $i = l+1, \ldots, m$, it holds that

$$r(y, \nu) = o(r(x, \mu)). \tag{31}$$

Proof Note that for (x, μ) close enough to $(\bar{x}, \bar{\mu})$, as $\alpha \in (0, 1)$, it holds that

$$(r(x, \mu))^{\alpha} \geq c_e r(x, \mu). \tag{32}$$

As already argued above, $\mathcal{A}(x, \mu) = \mathcal{A} = \mathcal{I}$. And, by Proposition 4, we have that (21) holds. In particular, by (6), by the hypothesis of the proposition, and by (21),

$$\|x - \bar{x}\| = O(r(x, \mu)), \quad \|\hat{x} - x\| = O(r(x, \mu)), \quad \|y - x\| = O(r(x, \mu)). \tag{33}$$

Hence, for $i \notin \mathcal{A} = \mathcal{I}$, by the continuity considerations it is easily seen that

$$r(x, \mu)\mu_i + \max\{g_i(\hat{x}) + \langle g_i'(x), y - \hat{x} \rangle, g_i(y)\} < \tfrac{1}{2} g_i(\bar{x}) < 0. \tag{34}$$

Next, using the definitions of $\mathcal{A}_1$ and of $Q_{\mathcal{A}}$, (32) and (21), for all $i \in \{l + 1, \ldots, m\} \cap \mathcal{A}_1$ we have that

$$\begin{aligned} \mu_i &\geq (r(x, \mu))^{\alpha} \geq c_e r(x, \mu) \geq \left| \tfrac{1}{r} g_i(\hat{x}) + \tfrac{1}{r} \langle g_i'(x), y - \hat{x} \rangle \right| \\ &\geq -\left(\tfrac{1}{r} g_i(\hat{x}) + \tfrac{1}{r} \langle g_i'(x), y - \hat{x} \rangle \right). \end{aligned}$$

Then, by the definition of ν, we conclude that for all $i \in \{l + 1, \ldots, m\} \cap \mathcal{A}_1(x, \mu)$,

$$\nu_i = \mu_i + \tfrac{1}{r} g_i(\hat{x}) + \tfrac{1}{r} \langle g_i'(x), y - \hat{x} \rangle. \tag{35}$$

By (34), $\nu_i = 0$ for $i \notin \mathcal{I}$. Then, by (20), it holds that

$$|\nu_i - \mu_i| = |\mu_i| = O(r(x, \mu)) \text{ for } i \notin \mathcal{I}.$$

Combining with (21), we obtain that

$$\|\nu - \mu\| = O(r(x, \mu)). \tag{36}$$

Next, since $g_i,\ i = 1, \ldots, m$, are twice continuously differentiable,

$$g_i(\hat{x}) = g_i(x) + \langle g_i'(x), \hat{x} - x\rangle + \tfrac{1}{2}\langle g_i''(x)(\hat{x} - x), \hat{x} - x\rangle + o(\|\hat{x} - x\|^2),$$
$$g_i(y) = g_i(x) + \langle g_i'(x), y - x\rangle + \tfrac{1}{2}\langle g_i''(x)(y - x), y - x\rangle + o(\|y - x\|^2).$$

Then, for all $i \in \{1, \ldots, m\}$ it holds that

$$\begin{aligned} g_i(\hat{x}) + \langle g_i'(x), y - \hat{x}\rangle - g_i(y) = \tfrac{1}{2}\langle g_i''(x)(\hat{x} - y), \hat{x} - y\rangle + o(\|\hat{x} - x\|^2) \\ + \langle g_i''(x)(\hat{x} - y), y - x\rangle + o(\|y - x\|^2). \end{aligned} \tag{37}$$

Thus, using (33), we obtain that

$$\begin{aligned} \|\nu - \lambda\| &\le \frac{1}{r}\|g(\hat{x}) + g'(x)(y - \hat{x}) - g(y)\| \\ &= \frac{1}{r}O(r^2) = O(r(x, \mu)). \end{aligned}$$

Hence, for the relation in (30), combining (33), (36) and the previous relation, we conclude that

$$\|y - x\| + \|\lambda - \mu\| = \|y - x\| + \|\lambda - \nu\| + \|\nu - \mu\| = O(r(x, \mu)).$$

Finally, we show that (31) holds. Note that, for $i \in \{1, \ldots, l\}$ we have

$$0 = g_i(\hat{x}) + \langle g_i'(x), y - \hat{x}\rangle - r(x, \mu)(\nu_i - \mu_i) = g_i(y) + s_i, \tag{38}$$

where, by (33), (36) and (37),

$$\begin{aligned} s_i &= g_i(\hat{x}) + \langle g_i'(x), y - \hat{x}\rangle - g_i(y) - r(x, \mu)(\nu_i - \mu_i) \\ &= O((r(x, \mu))^2) + O(r(x, \mu)\|\nu - \mu\|) = o(r(x, \mu)). \end{aligned}$$

Analogously, for $i \in \{l + 1, \ldots, m\}$, define

$$t_i = g_i(\hat{x}) + \langle g_i'(x), y - \hat{x}\rangle - g_i(y) - r(x, \mu)(\nu_i - \mu_i).$$

Then, by (33), (36) and (37), $t_i = o(r(x, \mu))$. Moreover, by the definition of ν_i it holds that

$$0 \le \nu_i, \quad g_i(y) + t_i \le 0, \quad \nu_i(g_i(y) + t_i) = 0, \tag{39}$$

for all $i \in \{l + 1, \ldots, m\}$.

Now, since $\nu_i = 0$ for $i \notin \mathcal{I}$, we conclude that

$$\begin{aligned} 0 = \phi'(y; x, \mu, \hat{x}, \hat{\mu}) &= \frac{\partial L}{\partial x}(x, \nu) + \frac{\partial^2 L}{\partial x^2}(x, \hat{\mu})(y - x) \\ &= \frac{\partial L}{\partial x}(y, \nu) + q, \end{aligned} \tag{40}$$

where, by the hypothesis that $\hat{\mu} \to \bar{\mu}$, by (36) and (33), it holds that

$$\begin{aligned} q &= \frac{\partial L}{\partial x}(x, \nu) + \frac{\partial^2 L}{\partial x^2}(x, \hat{\mu})(y - x) - \frac{\partial L}{\partial x}(y, \nu) \\ &= \left(\frac{\partial^2 L}{\partial x^2}(x, \hat{\mu}) - \frac{\partial^2 L}{\partial x^2}(x, \nu) \right) (y - x) + o(\|y - x\|) \\ &= o(r(x, \mu)). \end{aligned}$$

Combining (38)–(40), we obtain that (y, ν) solves the perturbed KKT system (28) for $p = (q, s, t) = o(r(x, \mu))$. Then, by the estimate (29), we conclude that

$$\|y - \bar{x}\| + \mathrm{dist}(\nu, \mathcal{M}(\bar{x})) \leq \tau \|p\| = o(r(x, \mu)).$$

Further, defining $\check{\nu}$ to be the projection of ν onto $\mathcal{M}(\bar{x})$, from the fact that r is locally Lipschitz-continuous and $r(\bar{x}, \check{\nu}) = 0$, from the latter estimate we obtain that

$$r(y, \nu) = r(y, \nu) - r(\bar{x}, \check{\nu}) = o(r(x, \mu)).$$

□

Our final result puts all the pieces together. In particular, it shows that solving the two QPs in Algorithm 1 provides an appropriate approximate solution to the augmented Lagrangian subproblem, satisfying conditions (9) in Theorem 1.

Theorem 4 *Let $(\bar{x}, \bar{\mu})$ satisfy SOSC (4).*

Then there exists a neighborhood of $(\bar{x}, \bar{\mu})$ such that for any (x, μ) in this neighborhood satisfying $r(x, \mu) > 0$ with $\mu_i \geq 0$ for $i = l + 1, \ldots, m$, it holds that

$$\frac{\partial \bar{L}}{\partial x}(x^+, \mu; r) = o(r(x, \mu)) \quad \text{and} \quad \|x^+ - x\| + \|\mu^+ - \mu\| = O(r(x, \mu)), \qquad (41)$$

where (x^+, μ^+) is such that $0 = \phi'(x^+; x, \mu, \tilde{x}, \tilde{\mu})$, $\mu_i^+ = \mu_i + \frac{1}{r} g_i(x^+)$ for $i = 1, \ldots, l$ and $\mu_i^+ = \max\{0, \mu_i + \frac{1}{r} g_i(x^+)\}$ for $i = l + 1, \ldots, m$, with $r = r(x, \mu)$; and $(\tilde{x}, \tilde{\mu})$ is such that $0 = \phi'(\tilde{x}; x, \mu, x, \mu)$, $\tilde{\mu}_i = \mu_i + \frac{1}{r} g_i(\tilde{x})$ for $i = 1, \ldots, l$ and $\tilde{\mu}_i = \max\{0, \mu_i + \frac{1}{r} g_i(\tilde{x})\}$ for $i = l + 1, \ldots, m$.

Consequently, the iterates generated by Algorithm 1 satisfy conditions (9) in Theorem 1, and thus any sequence $\{(x^k, \mu^k)\}$ generated by Algorithm 1 converges Q-superlinearly to $(\bar{x}, \hat{\mu})$ with some $\hat{\mu} \in \mathcal{M}(\bar{x})$.

Proof Using Proposition 5 with $\hat{x} = x$ and $\hat{\mu} = \mu$, we obtain for $\tilde{\nu} = \nu$ that

$$\|\tilde{x} - x\| + \|\tilde{\mu} - \mu\| = O(r(x, \mu)), \qquad r(\tilde{x}, \tilde{\nu}) = o(r(x, \mu)). \qquad (42)$$

Then, using again Proposition 5 with $\hat{x} = \tilde{x}$ and $\hat{\mu} = \tilde{\mu}$, we conclude that for $\nu^+ = \nu$

$$\|x^+ - x\| + \|\mu^+ - \mu\| = O(r(x, \mu)), \qquad r(x^+, \nu^+) = o(r(x, \mu)). \qquad (43)$$

To establish the first relation in (41), note that from (3) and (6) we have

$$\|\tilde{x} - \bar{x}\| = O(r(\tilde{x}, \tilde{\nu})), \ \|x^+ - \bar{x}\| = O(r(x^+, \nu^+)),$$
$$\left\| \frac{\partial L}{\partial x}(x^+, \nu^+) \right\| = O(r(x^+, \nu^+)).$$

Thus,

$$\tilde{x} - \bar{x} = o(r(x, \mu)), \quad x^+ - \bar{x} = o(r(x, \mu)), \quad \frac{\partial L}{\partial x}(x^+, \nu^+) = o(r(x, \mu)).$$

Since $x^+ - \tilde{x} = x^+ - \bar{x} + \bar{x} - \tilde{x} = o(r(x, \mu))$, from (37) with $\hat{x} = \tilde{x}$ and $y = x^+$ (noting that with these choices, since $x^+ - \tilde{x} = o(r)$, the second-order term on the right-hand side of (37) is of order $O(r)o(r)$), we conclude that

$$\begin{aligned} \|\nu^+ - \mu^+\| &\le \frac{1}{r}\|g(\tilde{x}) + g'(x)(x^+ - \tilde{x}) - g(x^+)\| \\ &= \frac{1}{r} o(r^2) = o(r(x, \mu)). \end{aligned}$$

Then, by the definition of the augmented Lagrangian $\bar{L}$,

$$\begin{aligned} \frac{\partial \bar{L}}{\partial x}(x^+, \mu; \sigma) &= \frac{\partial L}{\partial x}(x^+, \mu^+) \\ &= \frac{\partial L}{\partial x}(x^+, \nu^+) + (g'(x))^\top (\mu^+ - \nu^+) \\ &= o(r(x, \mu)). \end{aligned}$$

The equality above and the first relation in (43) mean that (41) is established. This corresponds to (9) in Theorem 1. The conclusions follow. □

4 Concluding Remarks

We have shown that locally, the cost of solving approximately the augmented Lagrangian subproblems in constrained optimization can be reduced to just two Newtonian inner iterations (two quadratic programming problems in the general case, or two linear systems when there are equality constraints only). Moreover, those two steps are sufficient for the overall fast convergence of the augmented Lagrangian algorithm, under the weakest assumptions (of second-order sufficiency, without any constraint qualifications and without strict complementarity). We conjecture that in the equality-constrained case, it might be posssible to relax second-order sufficiency to the weaker assumption of noncriticality of the Lagrange multiplier associated to the solution.

Acknowledgements We thank the three anonymous referees for their evaluations of the paper, which led to an improved version. We are specially grateful to the referee who pointed out one technical issue that required a correction.

References

1. Andreani, R., Birgin, E.G., Martínez, J.M., Schuverdt, M.L.: Augmented Lagrangian methods under the constant positive linear dependence constraint qualification. Math. Program. **111**(1–2, Ser. B), 5–32 (2008)
2. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Computer Science and Applied Mathematics. Academic Press Inc, New York (1982)
3. Birgin, E.G., Martínez, J.M.: Improving ultimate convergence of an augmented Lagrangian method. Optim. Methods Softw. **23**(2), 177–195 (2008)
4. Birgin, E.G., Martínez, J.M.: Practical Augmented Lagrangian Methods for Constrained Optimization, Fundamentals of Algorithms, vol. 10. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2014)
5. Daryina, A.N., Izmailov, A.F., Solodov, M.V.: A class of active-set Newton methods for mixed complementarity problems. SIAM J. Optim. **15**(2), 409–429 (2004/2005)
6. Facchinei, F., Fischer, A., Kanzow, C.: On the accurate identification of active constraints. SIAM J. Optim. **9**(1), 14–32 (1999)
7. Fernández, D., Pilotta, E.A., Torres, G.A.: An inexact restoration strategy for the globalization of the sSQP method. Comput. Optim. Appl. **54**(3), 595–617 (2013)
8. Fernández, D., Solodov, M.: Stabilized sequential quadratic programming for optimization and a stabilized Newton-type method for variational problems. Math. Program. **125**(1, Ser. A), 47–73 (2010)
9. 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**(2), 384–407 (2012)
10. Fischer, A.: Local behavior of an iterative framework for generalized equations with nonisolated solutions. Math. Program. **94**(1, Ser. A), 91–124 (2002)
11. Gill, P.E., Kungurtsev, V., Robinson, D.P.: A stabilized SQP method: global convergence. IMA J. Numer. Anal. **37**(1), 407–443 (2017)
12. Gill, P.E., Kungurtsev, V., Robinson, D.P.: A stabilized SQP method: superlinear convergence. Math. Program. **163**(1–2, Ser. A), 369–410 (2017)
13. Hager, W.W.: Stabilized sequential quadratic programming. Comput. Optim. Appl. **12**(1–3), 253–273 (1999)
14. Hager, W.W., Gowda, M.S.: Stability in the presence of degeneracy and error estimation. Math. Program. **85**(1, Ser. A), 181–192 (1999)
15. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**, 303–320 (1969)
16. 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**(1–2, Ser. A), 591–604 (2013)
17. Izmailov, A.F., Kurennoy, A.S., Solodov, M.V.: Local convergence of the method of multipliers for variational and optimization problems under the noncriticality assumption. Comput. Optim. Appl. **60**(1), 111–140 (2015)
18. Izmailov, A.F., Solodov, M.V.: Newton-Type Methods for Optimization and Variational Problems. Springer Series in Operations Research and Financial Engineering. Springer, Cham (2014)
19. Izmailov, A.F., Solodov, M.V., Uskov, E.I.: Combining stabilized SQP with the augmented Lagrangian algorithm. Comput. Optim. Appl. **62**(2), 405–429 (2015)
20. Izmailov, A.F., Solodov, M.V., Uskov, E.I.: Globalizing stabilized sequential quadratic programming method by smooth primal–dual exact penalty function. J. Optim. Theory Appl. **169**(1), 148–178 (2016)
21. Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In: Optimization (Symposium, University of Keele, Keele, 1968), pp. 283–298. Academic Press, London (1969)
22. Rockafellar, R.T.: A dual approach to solving nonlinear programming problems by unconstrained optimization. Math. Program. **5**, 354–373 (1973)
23. Wright, S.J.: An algorithm for degenerate nonlinear programming with rapid local convergence. SIAM J. Optim. **15**(3), 673–696 (2005)