SCIENCE CHINA
Mathematics

• ARTICLES •

August 2016 Vol. 59 No. 8: 1525–1542
doi: 10.1007/s11425-016-0300-6

An active set algorithm for nonlinear optimization with polyhedral constraints

HAGER William W.[1] & ZHANG Hongchao[2,*]

[1]*Department of Mathematics, University of Florida, Gainesville, FL* 32611-8105*, USA;*
[2]*Department of Mathematics, Louisiana State University, Baton Rouge, LA* 70803-4918*, USA*
Email: hager@ufl.edu, hozhang@math.lsu.edu

Received November 2, 2015; accepted May 3, 2016; published online June 1, 2016

Abstract A polyhedral active set algorithm PASA is developed for solving a nonlinear optimization problem whose feasible set is a polyhedron. Phase one of the algorithm is the gradient projection method, while phase two is any algorithm for solving a linearly constrained optimization problem. Rules are provided for branching between the two phases. Global convergence to a stationary point is established, while asymptotically PASA performs only phase two when either a nondegeneracy assumption holds, or the active constraints are linearly independent and a strong second-order sufficient optimality condition holds.

Keywords polyhedral constrained optimization, active set algorithm, PASA, gradient projection algorithm, local and global convergence

MSC(2010) 90C06, 90C26, 65Y20

Citation: Hager W W, Zhang H. An active set algorithm for nonlinear optimization with polyhedral constraints. Sci China Math, 2016, 59: 1525–1542, doi: 10.1007/s11425-016-0300-6

1 Introduction

We develop an active set algorithm for a general nonlinear polyhedral constrained optimization problem

$$\min\{f(\boldsymbol{x}) : \boldsymbol{x} \in \Omega\}, \quad \text{where } \Omega = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} \leqslant \boldsymbol{b}\}. \tag{1.1}$$

Here f is a real-valued, continuously differentiable function, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, and Ω is assumed to be nonempty. In an earlier paper [27], we developed an active set algorithm for bound constrained optimization. In this paper, we develop a new machinery for handling the more complex polyhedral constraints of (1.1). Our polyhedral active set algorithm (PASA) has two phases: Phase one is the gradient projection algorithm, while phase two is any algorithm for solving a linearly constrained optimization problem over a face of the polyhedron Ω. The gradient projection algorithm of phase one is robust in the sense that it converges to a stationary point under mild assumptions, but the convergence rate is often linear at best. When optimizing over a face of the polyhedron in phase two, we could accelerate the convergence through the use of a superlinearly convergent algorithm based on conjugate gradients, a quasi-Newton update, or a Newton iteration. In this paper, we give rules for switching between phases which ensure that asymptotically, only phase two is performed. Hence, the asymptotic convergence rate of PASA coincides with the convergence rate of the scheme used to solve the linearly constrained problem of phase two. A separate paper will focus on a specific numerical implementation of PASA.

*Corresponding author

We briefly survey some of the rich history of active set methods. Some of the initial work focused on the use of the conjugate gradient method with bound constraints as in [13, 15–17, 38, 39, 47]. Work on gradient projection methods includes [1, 7, 25, 35, 37, 42]. Convergence is accelerated by using Newton and trust region methods [12]. Superlinear and quadratic convergence for nondegenerate problems can be found in [2, 6, 11, 18], while analogous convergence results are given in [19, 21, 34, 36], even for degenerate problems. The affine scaling interior point approach [3, 8–10, 14, 28, 30, 33, 43, 49] is related to the trust region algorithm. Linear, superlinear, and quadratic convergence results have been established.

Recent developments on active set methods for quadratic programming problem can be found in [20,23]. A treatment of active set methods in a rather general setting is given in [32]. We also point out the recent work [24] on a very efficient two-phase active set method for conic-constrained quadratic programming, and the earlier work [22] on a two-phase active set method for quadratic programming. As in [27], the first phase in both applications is the gradient projection method. The second phase is a Newton method in [24], while it is a linear solver in [22]. Note that PASA applies to the general nonlinear objective in (1.1). Active set strategies were applied to ℓ_1 minimization in [44, 45] and in [29] they were applied to the minimization of a nonsmooth dual problem that arises when projecting a point onto a polyhedron. In [44, 45], a nonmonotone line search based on "Shrinkage" is used to estimate a support at the solution, while a nonmonotone SpaRSA algorithm [26, 46] is used in [29] to approximately identify active constraints.

Unlike most active set methods in the literature, our algorithm is not guaranteed to identify the active constraints in a finite number of iterations due to the structure of the line search in the gradient projection phase. Instead, we show that only the fast phase two algorithm is performed asymptotically, even when strict complementary slackness is violated. Moreover, our line search only requires one projection in each iteration, while algorithms that identify active constraints often employ a piecewise projection scheme that may require additional projections when the stepsize increases.

The paper is organized as follows. Section 2 gives a detailed statement of the polyhedral active set algorithm, while Section 3 establishes its global convergence. Section 4 gives some properties for the solution and multipliers associated with a Euclidean projection onto Ω. Section 5 shows that asymptotically PASA performs only phase two when converging to a nondegenerate stationary point, while Section 6 establishes the analogous result for degenerate problems when the active constraint gradients are linearly independent and a strong second-order sufficient optimality condition holds. Finally, Section 7 concludes the paper.

Notation. Throughout the paper, c denotes a generic nonnegative constant which has different values in different inequalities. For any set $\mathcal{S}$, $|\mathcal{S}|$ stands for the number of elements (cardinality) of $\mathcal{S}$, while $\mathcal{S}^c$ is the complement of $\mathcal{S}$. The set $\mathcal{S}-\boldsymbol{x}$ is defined by $\mathcal{S}-\boldsymbol{x}=\{\boldsymbol{y}-\boldsymbol{x}:\boldsymbol{y}\in\mathcal{S}\}$. The distance between a set $\mathcal{S}\subset\mathbb{R}^n$ and a point $\boldsymbol{x}\in\mathbb{R}^n$ is given by

$$\operatorname{dist}(\boldsymbol{x},\mathcal{S})=\inf\{\|\boldsymbol{x}-\boldsymbol{y}\|:\boldsymbol{y}\in\mathcal{S}\},$$

where $\|\cdot\|$ is the Euclidean norm. The subscript k is often used to denote the iteration number in an algorithm, while x_{ki} stands for the i-th component of the iterate $\boldsymbol{x}_k$. The gradient $\nabla f(\boldsymbol{x})$ is a row vector while $\boldsymbol{g}(\boldsymbol{x})=\nabla f(\boldsymbol{x})^{\mathrm{T}}$ is the gradient arranged as a column vector; here $^{\mathrm{T}}$ denotes transpose. The gradient at the iterate $\boldsymbol{x}_k$ is $\boldsymbol{g}_k=\boldsymbol{g}(\boldsymbol{x}_k)$. In several theorems, we assume that f is Lipschitz continuously differentiable in a neighborhood of a stationary point $\boldsymbol{x}^*$. The Lipschitz constant for ∇f is always denoted κ. We let $\nabla^2 f(\boldsymbol{x})$ denote the Hessian of f at $\boldsymbol{x}$. The ball with center $\boldsymbol{x}$ and radius r is denoted $\mathcal{B}_r(\boldsymbol{x})$. For any matrix $\boldsymbol{M}$, $\mathcal{N}(\boldsymbol{M})$ is the null space. If $\mathcal{S}$ is a subset of the row indices of $\boldsymbol{M}$, then $\boldsymbol{M}_{\mathcal{S}}$ denotes the submatrix of $\boldsymbol{M}$ with row indices $\mathcal{S}$. For any vector $\boldsymbol{b}$, $\boldsymbol{b}_{\mathcal{S}}$ is the subvector of $\boldsymbol{b}$ with indices $\mathcal{S}$. $P_\Omega(\boldsymbol{x})$ denotes the Euclidean projection of $\boldsymbol{x}$ onto Ω:

$$\mathcal{P}_\Omega(\boldsymbol{x})=\arg\min\{\|\boldsymbol{x}-\boldsymbol{y}\|:\boldsymbol{y}\in\Omega\}. \tag{1.2}$$

For any $\boldsymbol{x}\in\Omega$, the active and free index sets are defined by $\mathcal{A}(\boldsymbol{x})=\{i:(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})_i=0\}$ and $\mathcal{F}(\boldsymbol{x})=\{i:(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})_i<0\}$, respectively.

2 Structure of the algorithm

As explained in the introduction, PASA uses the gradient projection algorithm in phase one and a linearly constrained optimization algorithm in phase two. Algorithm 1 is the gradient projection algorithm (GPA) used for the analysis in this paper. A cartoon of the algorithm appears in Figure 1.

Algorithm 1: Prototype gradient projection algorithm (GPA).

Parameters: δ and $\eta \in (0,1)$, $\alpha \in (0,\infty)$

While stopping condition does not hold

1. $\boldsymbol{d}_k = \boldsymbol{y}(\boldsymbol{x}_k,\alpha) - \boldsymbol{x}_k$, $\boldsymbol{y}(\boldsymbol{x},\alpha) = \mathcal{P}_\Omega(\boldsymbol{x} - \alpha \boldsymbol{g}(\boldsymbol{x}))$
2. $s_k = \eta^j$ where $j \geqslant 0$ is smallest integer such that $f(\boldsymbol{x}_k + s_k\boldsymbol{d}_k) \leqslant f(\boldsymbol{x}_k) + s_k\delta\nabla f(\boldsymbol{x}_k)\boldsymbol{d}_k$
3. $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + s_k\boldsymbol{d}_k$ and $k \leftarrow k+1$

End while

This is a simple monotone algorithm based on an Armijo line search. Better numerical performance is achieved with a more general nonmonotone line search such as that given in [27], and all the analysis directly extends to this more general framework; however, to simplify the analysis and discussion in the paper, we utilize Algorithm 1 for the GPA.

The requirements for the linearly constrained optimizer (LCO) of phase two, which operates on the *faces* of Ω, are now developed. One of the requirements is that when the active sets repeat in an infinite series of iterations. Then the iterates must approach stationary. To formulate this requirement in a precise way, we define

$$\boldsymbol{g}^{\mathcal{I}}(\boldsymbol{x}) = \mathcal{P}_{\mathcal{N}(\boldsymbol{A}_{\mathcal{I}})}(\boldsymbol{g}(\boldsymbol{x})) = \arg\min\{\|\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{x})\| : \boldsymbol{y} \in \mathbb{R}^n \text{ and } \boldsymbol{A}_{\mathcal{I}}\boldsymbol{y} = \boldsymbol{0}\}. \tag{2.1}$$

Thus $\boldsymbol{g}^{\mathcal{I}}(\boldsymbol{x})$ is the projection of the gradient $\boldsymbol{g}(\boldsymbol{x})$ onto the null space $\mathcal{N}(\boldsymbol{A}_{\mathcal{I}})$. We also let $\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x})$ denote

$$\boldsymbol{g}^{\mathcal{I}}(\boldsymbol{x}) \quad \text{for} \quad \mathcal{I} = \mathcal{A}(\boldsymbol{x}).$$

If $\mathcal{A}(\boldsymbol{x})$ is empty, then $\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x})$, while if $\boldsymbol{x}$ is a vertex of Ω, then $\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}) = \boldsymbol{0}$. This suggests that $e(\boldsymbol{x}) = \|\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x})\|$ represents a local measure of stationarity in the sense that it vanishes if and only if $\boldsymbol{x}$ is a stationary point on its associated face

$$\{\boldsymbol{y} \in \Omega : (\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b})_i = 0 \text{ for all } i \in \mathcal{A}(\boldsymbol{x})\}.$$

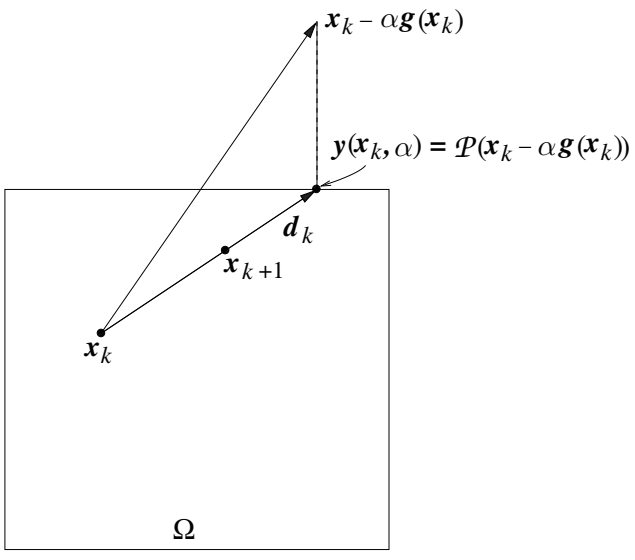

Figure 1 An iteration of the gradient projection algorithm

The requirements for the phase two LCO are the following:

F1. $\boldsymbol{x}_k \in \Omega$ and $f(\boldsymbol{x}_{k+1}) \leqslant f(\boldsymbol{x}_k)$ for each k.

F2. $\mathcal{A}(\boldsymbol{x}_k) \subset \mathcal{A}(\boldsymbol{x}_{k+1})$ for each k.

F3. If $\mathcal{A}(\boldsymbol{x}_{j+1}) = \mathcal{A}(\boldsymbol{x}_j)$ for $j \geqslant k$, then $\liminf_{j\to\infty} e(\boldsymbol{x}_j) = 0$.

Condition F1 requires that the iterates in phase two are monotone, in contrast to phase one where the iterates could be nonmonotone. By F2, the active set only grows during phase two, while F3 implies that the local stationarity measure becomes small when the active set does not change. Conditions F1–F3 are easily fulfilled by algorithms based on gradient or Newton type iterations which employ a monotone line search and which add constraints to the active set whenever a new constraint becomes active.

Our decision for switching between phase one (GPA) and phase two (LCO) is based on a comparison of two different measures of stationarity. One measure is the local stationarity measure $e(\cdot)$, introduced already, which measures stationarity relative to a face of Ω. The second measure of stationarity is a global metric in the sense that it vanishes at $\boldsymbol{x}$ if and only if $\boldsymbol{x}$ is a stationary point for the optimization problem (1.1). For $\alpha \geqslant 0$, let $\boldsymbol{y}(\boldsymbol{x}, \alpha)$ be the point obtained by taking a step from $\boldsymbol{x}$ along the negative gradient and projecting onto Ω, i.e.,

$$\boldsymbol{y}(\boldsymbol{x}, \alpha) = \mathcal{P}_\Omega(\boldsymbol{x} - \alpha \boldsymbol{g}(\boldsymbol{x})) = \arg\min \left\{ \frac{1}{2}\|\boldsymbol{x} - \alpha \boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{y}\|^2 : \boldsymbol{A}\boldsymbol{y} \leqslant \boldsymbol{b} \right\}. \tag{2.2}$$

The vector

$$\boldsymbol{d}^\alpha(\boldsymbol{x}) = \boldsymbol{y}(\boldsymbol{x}, \alpha) - \boldsymbol{x} \tag{2.3}$$

points from $\boldsymbol{x}$ to the projection of $\boldsymbol{x} - \alpha \boldsymbol{g}(\boldsymbol{x})$ onto Ω. As seen in [27, Proposition 2.1], when $\alpha > 0$, $\boldsymbol{d}^\alpha(\boldsymbol{x}) = \boldsymbol{0}$ if and only if $\boldsymbol{x}$ is a stationary point for (1.1). We monitor convergence to a stationary point using the function E defined by

$$E(\boldsymbol{x}) = \|\boldsymbol{d}^1(\boldsymbol{x})\|.$$

$E(\boldsymbol{x})$ vanishes if and only if $\boldsymbol{x}$ is a stationary point of (1.1). If $\Omega = \mathbb{R}^n$, then $E(\boldsymbol{x}) = \|\boldsymbol{g}(\boldsymbol{x})\|$, the norm of the gradient, which is the usual way to assess convergence to a stationary point in unconstrained optimization.

The rules for switching between phase one and phase two depend on the relative size of the stationarity measures E and e. We choose a parameter $\theta \in (0, 1)$ and branch from phase one to phase two when $e(\boldsymbol{x}_k) \geqslant \theta E(\boldsymbol{x}_k)$. Similarly, we branch from phase two to phase one when $e(\boldsymbol{x}_k) < \theta E(\boldsymbol{x}_k)$. To ensure that only phase two is executed asymptotically at a degenerate stationary point, we may need to decrease θ as the iterates converge. The decision to decrease θ is based on what we called the undecided index set $\mathcal{U}$ which is defined as follows. Let $\boldsymbol{x}$ denote the current iterate, let $E(\boldsymbol{x})$ be the global measure of stationarity, and let $\boldsymbol{\lambda}(\boldsymbol{x})$ denote any Lagrange multiplier associated with the polyhedral constraint in (2.2) and $\alpha = 1$, i.e., if $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}, 1)$ is the solution of (2.2) for $\alpha = 1$, then $\boldsymbol{\lambda}(\boldsymbol{x})$ is any vector that satisfies the conditions

$$\boldsymbol{y} - \boldsymbol{x} + \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{0}, \quad \boldsymbol{\lambda}(\boldsymbol{x}) \geqslant \boldsymbol{0}, \quad \lambda_i(\boldsymbol{x}) = 0 \quad \text{if} \quad i \in \mathcal{F}(\boldsymbol{y}).$$

Given parameters $\beta \in (1, 2)$ and $\gamma \in (0, 1)$, the undecided index set is defined by

$$\mathcal{U}(\boldsymbol{x}) = \{i : \lambda_i(\boldsymbol{x}) \geqslant E(\boldsymbol{x})^\gamma \text{ and } (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})_i \geqslant E(\boldsymbol{x})^\beta\}.$$

If $\boldsymbol{x}$ is close enough to a stationary point that $E(\boldsymbol{x})$ is small, then the indices in $\mathcal{U}(\boldsymbol{x})$ correspond to those constraints for which the associated multiplier $\lambda_i(\boldsymbol{x})$ is relatively large in the sense that $\lambda_i(\boldsymbol{x}) \geqslant E(\boldsymbol{x})^\gamma$, and the i-th constraint is relatively inactive in the sense that $(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{y})_i \geqslant E(\boldsymbol{x})^\beta$. By the first-order optimality conditions at a local minimizer, large multipliers are associated with active constraints. Hence, when the multiplier is relatively large and the constraint is relatively inactive, we consider the constraint undecided. When $\mathcal{U}(\boldsymbol{x})$ is empty, then we feel that the active constraints are nearly identified, so we decrease θ in phase one so that phase two will compute a more accurate local stationary point before branching back to phase one.

Algorithm 2 is the polyhedral active set algorithm (PASA). The parameter ϵ is the convergence tolerance, the parameter θ controls the branching between phase one and phase two, while the parameter μ controls the decay of θ when the undecided index set is empty.

Algorithm 2: Polyhedral active set algorithm (PASA).

```
Parameters:  ε ∈ [0, ∞), θ and μ ∈ (0, 1)
x_1 = P_Ω(x_0), k = 1
Phase one:  While E(x_k) > ε execute GPA
  If U(x_k) = ∅ and e(x_k) < θE(x_k), then θ ← μθ.
  If e(x_k) ⩾ θE(x_k), goto phase two.
End while
Phase two:  While E(x_k) > ε execute LCO
  If e(x_k) < θE(x_k), goto phase one.
End while
```

3 Global convergence

Since Algorithm 1, GPA, is a special case of the nonmonotone gradient projection algorithm studied in [27], our previously established global convergence result, stated below, holds.

Theorem 3.1. *Let $\mathcal{L}$ be the level set defined by*

$$\mathcal{L} = \{\boldsymbol{x} \in \Omega : f(\boldsymbol{x}) \leqslant f(\boldsymbol{x}_0)\}. \tag{3.1}$$

Assume the following conditions hold:

A1. *f is bounded from below on $\mathcal{L}$ and $d_{\max} = \sup_k \|\boldsymbol{d}_k\| < \infty$.*

A2. *If $\bar{\mathcal{L}}$ is the collection of $\boldsymbol{x} \in \Omega$ whose distance to $\mathcal{L}$ is at most $d_{\max}$, then ∇f is Lipschitz continuous on $\bar{\mathcal{L}}$.*

Then GPA with $\epsilon = 0$ either terminates in a finite number of iterations at a stationary point, or we have

$$\liminf_{k\to\infty} E(\boldsymbol{x}_k) = 0.$$

The global convergence of PASA essentially follows from the global convergence of GPA and the requirement F3 for the linearly constrained optimizer.

Theorem 3.2. *If the assumptions of Theorem* 3.1 *hold and the linearly constrained optimizer satisfies* F1–F3*, then PASA with $\epsilon = 0$ either terminates in a finite number of iterations at a stationary point, or we have*

$$\liminf_{k\to\infty} E(\boldsymbol{x}_k) = 0. \tag{3.2}$$

Proof. If only phase one is performed for k sufficiently large, then (3.2) follows from Theorem 3.1. If only phase two is performed for k sufficiently large, then $e(\boldsymbol{x}_k) \geqslant \theta E(\boldsymbol{x}_k)$ for k sufficiently large. Since θ is only changed in phase one, we can treat θ as a fixed positive scalar for k sufficiently large. By F2, the active sets approach a fixed limit for k sufficiently large. By F3 and the inequality $e(\boldsymbol{x}_k) \geqslant \theta E(\boldsymbol{x}_k)$, (3.2) holds. Finally, suppose that there are an infinite number of branches from phase two to phase one. If (3.2) does not hold, then there exists $\tau > 0$ such that $E(\boldsymbol{x}_k) = \|\boldsymbol{d}^1(\boldsymbol{x}_k)\| \geqslant \tau$ for all k. By [27, Property P6] and the definition $\boldsymbol{d}_k = \boldsymbol{d}^\alpha(\boldsymbol{x}_k)$, we have

$$\nabla f(\boldsymbol{x}_k)\boldsymbol{d}_k = \boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}^\alpha(\boldsymbol{x}_k) \leqslant -\|\boldsymbol{d}^\alpha(\boldsymbol{x}_k)\|^2/\alpha = -\|\boldsymbol{d}_k\|^2/\alpha, \tag{3.3}$$

which implies that

$$\frac{|\boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k|}{\|\boldsymbol{d}_k\|^2} \geqslant \frac{1}{\alpha}. \tag{3.4}$$

By [27, Properties P4 and P5] and the lower bound $\|\boldsymbol{d}^1(\boldsymbol{x}_k)\| \geqslant \tau$, we have

$$\|\boldsymbol{d}_k\| = \|\boldsymbol{d}^\alpha(\boldsymbol{x}_k)\| \geqslant \min\{\alpha, 1\}\|\boldsymbol{d}^1(\boldsymbol{x}_k)\| \geqslant \min\{\alpha, 1\}\tau. \tag{3.5}$$

For the Armijo line search in GPA, it follows from [48, Lemma 2.1] that

$$s_k \geqslant \min\left\{1, \left(\frac{2\eta(1-\delta)}{\kappa}\right)\frac{|\boldsymbol{g}_k^{\mathrm{T}}\boldsymbol{d}_k|}{\|\boldsymbol{d}_k\|^2}\right\}$$

for all iterations in GPA, where κ is the Lipschitz constant for ∇f. Combine this with (3.4) to obtain

$$s_k \geqslant \min\left\{1, \left(\frac{2\eta(1-\delta)}{\kappa\alpha}\right)\right\}. \tag{3.6}$$

By the line search condition in Step 2 of GPA, it follows from (3.3), (3.5) and (3.6) that there exists $c > 0$ such that

$$f(\boldsymbol{x}_{k+1}) \leqslant f(\boldsymbol{x}_k) - c \tag{3.7}$$

for all iterations in GPA with k sufficiently large. Since the objective function decreases monotonically in phase two, and since there are an infinite number of iterations in GPA, (3.7) contradicts Assumption A1 that f is bounded from below. □

4 Properties of projections

The proof that only phase two of PASA is executed asymptotically relies on some properties for the solution of the projection problem (2.2) that are established in this section. Since the projection onto a convex set is a nonexpansive operator, we have

$$\begin{aligned}\|\boldsymbol{y}(\boldsymbol{x}_1,\alpha) - \boldsymbol{y}(\boldsymbol{x}_2,\alpha)\| &= \|\mathcal{P}_\Omega(\boldsymbol{x}_1 - \alpha\boldsymbol{g}(\boldsymbol{x}_1)) - \mathcal{P}_\Omega(\boldsymbol{x}_2 - \alpha\boldsymbol{g}(\boldsymbol{x}_2))\| \\ &\leqslant \|(\boldsymbol{x}_1 - \alpha\boldsymbol{g}(\boldsymbol{x}_1)) - (\boldsymbol{x}_2 - \alpha\boldsymbol{g}(\boldsymbol{x}_2))\| \\ &\leqslant (1+\alpha\kappa)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|,\end{aligned} \tag{4.1}$$

where κ is a Lipschitz constant for $\boldsymbol{g}$. Since $\boldsymbol{d}^\alpha(\boldsymbol{x}^*) = \boldsymbol{0}$ for all $\alpha > 0$ when $\boldsymbol{x}^*$ is a stationary point, it follows that $\boldsymbol{y}(\boldsymbol{x}^*,\alpha) = \boldsymbol{x}^*$ for all $\alpha > 0$. In the special case where $\boldsymbol{x}_2 = \boldsymbol{x}^*$, (4.1) yields

$$\|\boldsymbol{y}(\boldsymbol{x},\alpha) - \boldsymbol{x}^*\| = \|\boldsymbol{y}(\boldsymbol{x},\alpha) - \boldsymbol{y}(\boldsymbol{x}^*,\alpha)\| \leqslant (1+\alpha\kappa)\|\boldsymbol{x} - \boldsymbol{x}^*\|. \tag{4.2}$$

Similar to (4.1), but with $\boldsymbol{y}$ replaced by $\boldsymbol{d}^\alpha$, we have

$$\|\boldsymbol{d}^\alpha(\boldsymbol{x}_1) - \boldsymbol{d}^\alpha(\boldsymbol{x}_2)\| \leqslant (2+\alpha\kappa)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|. \tag{4.3}$$

Next, let us develop some properties for the multipliers associated with the constraint in (2.2). The first-order optimality conditions associated with (2.2) can be expressed as follows: At any solution $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x},\alpha)$ of (2.2), there exists a multiplier $\boldsymbol{\lambda} \in \mathbb{R}^m$ such that

$$\boldsymbol{y} - \boldsymbol{x} + \alpha\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = \boldsymbol{0}, \quad \boldsymbol{\lambda} \geqslant \boldsymbol{0}, \quad \lambda_i = 0 \quad \text{if} \quad i \in \mathcal{F}(\boldsymbol{y}). \tag{4.4}$$

Let $\Lambda(\boldsymbol{x},\alpha)$ denote the set of multipliers $\boldsymbol{\lambda}$ satisfying (4.4) at the solution $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x},\alpha)$ of (2.2). If $\boldsymbol{x}^*$ is a stationary point for (1.1) and $\alpha > 0$, then $\boldsymbol{y}(\boldsymbol{x}^*,\alpha) = \boldsymbol{x}^*$, and the first equation in (4.4) reduces to

$$\boldsymbol{g}(\boldsymbol{x}^*) + \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{\lambda}/\alpha) = \boldsymbol{0},$$

which is the gradient of the Lagrangian for (1.1), but with the multiplier scaled by α. Since $\mathcal{F}(\boldsymbol{y}(\boldsymbol{x}^*,\alpha)) = \mathcal{F}(\boldsymbol{x}^*)$, $\boldsymbol{\lambda}/\alpha$ is a multiplier for the constraint in (1.1). Thus if $\boldsymbol{x}^*$ is a stationary point for (1.1) and $\Lambda(\boldsymbol{x}^*)$ is the set of Lagrange multipliers associated with the constraint, we have

$$\Lambda(\boldsymbol{x}^*,\alpha) = \alpha\Lambda(\boldsymbol{x}^*). \tag{4.5}$$

By (4.2) $\boldsymbol{y}(\boldsymbol{x},\alpha)$ approaches $\boldsymbol{x}^*$ as $\boldsymbol{x}$ approaches $\boldsymbol{x}^*$. Consequently, the indices $\mathcal{F}(\boldsymbol{x}^*)$ free at $\boldsymbol{x}^*$ are free at $\boldsymbol{y}(\boldsymbol{x},\alpha)$ when $\boldsymbol{x}$ is sufficiently close to $\boldsymbol{x}^*$. The multipliers associated with (2.2) have the following stability property.

Proposition 4.1. *Suppose $\boldsymbol{x}^*$ is a stationary point for (1.1) and for some $r > 0$, $\boldsymbol{g}$ is Lipschitz continuous in $\mathcal{B}_r(\boldsymbol{x}^*)$ with Lipschitz constant κ. If $\alpha \geqslant 0$ and $\boldsymbol{x} \in \mathcal{B}_r(\boldsymbol{x}^*)$ is close enough to $\boldsymbol{x}^*$ that $\mathcal{F}(\boldsymbol{x}^*) \subset \mathcal{F}(\boldsymbol{y}(\boldsymbol{x}, \alpha))$, then*

$$\operatorname{dist}\{\boldsymbol{\lambda}, \Lambda(\boldsymbol{x}^*, \alpha)\} \leqslant 2c(1 + \kappa\alpha)\|\boldsymbol{x} - \boldsymbol{x}^*\|$$

for all $\boldsymbol{\lambda} \in \Lambda(\boldsymbol{x}, \alpha)$, where c is independent of $\boldsymbol{x}$ and depends only on $\boldsymbol{A}$.

Proof. This is essentially a consequence of the upper Lipschitzian properties of polyhedral multifunctions as established in [40, Proposition 1] or [41, Corollary 4.2]. Here is a short proof based on Hoffman's stability result [31] for a perturbed linear system of inequalities. Since $\mathcal{F}(\boldsymbol{x}^*) \subset \mathcal{F}(\boldsymbol{y}(\boldsymbol{x}, \alpha))$, it follows that any $\boldsymbol{\lambda} \in \Lambda(\boldsymbol{x}, \alpha)$ is feasible in the system

$$\boldsymbol{p} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = \boldsymbol{0}, \quad \boldsymbol{\lambda} \geqslant \boldsymbol{0}, \quad \lambda_i = 0 \quad \text{if} \quad i \in \mathcal{F}(\boldsymbol{x}^*),$$

with $\boldsymbol{p} = \boldsymbol{p}_1 := \boldsymbol{y}(\boldsymbol{x}, \alpha) - \boldsymbol{x} + \alpha\boldsymbol{g}(\boldsymbol{x})$. Since $\boldsymbol{x}^*$ is a stationary point for (1.1), the elements of $\Lambda(\boldsymbol{x}^*, \alpha)$ are feasible in the same system but with $\boldsymbol{p} = \boldsymbol{p}_2 := \boldsymbol{x}^* - \boldsymbol{x}^* + \alpha\boldsymbol{g}(\boldsymbol{x}^*)$. Hence, by Hoffman's result [31], there exists a constant c, independent of $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ and depending only on $\boldsymbol{A}$, such that

$$\operatorname{dist}\{\boldsymbol{\lambda}, \Lambda(\boldsymbol{x}^*, \alpha)\} \leqslant c\|\boldsymbol{p}_1 - \boldsymbol{p}_2\|.$$

We use (4.2) to obtain

$$\|\boldsymbol{p}_1 - \boldsymbol{p}_2\| = \|(\boldsymbol{y}(\boldsymbol{x}, \alpha) - \boldsymbol{x}^*) + (\boldsymbol{x}^* - \boldsymbol{x}) + \alpha(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}^*))\| \leqslant 2(1 + \alpha\kappa)\|\boldsymbol{x} - \boldsymbol{x}^*\|,$$

which completes the proof. □

In the following proposition, we study the projection of a step $\boldsymbol{x}_k - \alpha\boldsymbol{g}(\boldsymbol{x}_k)$ onto the subset Ω_k of Ω that also satisfies the active constraints at $\boldsymbol{x}_k$. We show that the step can be replaced by $\boldsymbol{x}_k - \alpha\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)$ without effecting the projection.

Proposition 4.2. *For all $\alpha \geqslant 0$, we have*

$$\mathcal{P}_{\Omega_k}(\boldsymbol{x}_k - \alpha\boldsymbol{g}(\boldsymbol{x}_k)) = \mathcal{P}_{\Omega_k}(\boldsymbol{x}_k - \alpha\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)), \tag{4.6}$$

where

$$\Omega_k = \{\boldsymbol{x} \in \Omega : (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})_i = 0 \text{ for all } i \in \mathcal{A}(\boldsymbol{x}_k)\}. \tag{4.7}$$

Proof. Let $\boldsymbol{p}$ be defined by

$$\boldsymbol{p} = \mathcal{P}_{\Omega_k}(\boldsymbol{x}_k - \alpha\boldsymbol{g}(\boldsymbol{x}_k)) - \boldsymbol{x}_k = \arg\min\{\|\boldsymbol{x}_k - \alpha\boldsymbol{g}(\boldsymbol{x}_k) - \boldsymbol{y}\| : \boldsymbol{y} \in \Omega_k\} - \boldsymbol{x}_k. \tag{4.8}$$

With the change of variables $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{x}_k$, we can write

$$\boldsymbol{p} = \arg\min\{\|\boldsymbol{z} + \alpha\boldsymbol{g}(\boldsymbol{x}_k)\| : \boldsymbol{z} \in \Omega_k - \boldsymbol{x}_k\}. \tag{4.9}$$

Since $\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)$ is the orthogonal projection of $\boldsymbol{g}(\boldsymbol{x}_k)$ onto the null space $\mathcal{N}(\boldsymbol{A}_{\mathcal{I}})$, where $\mathcal{I} = \mathcal{A}(\boldsymbol{x}_k)$, the difference $\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k) - \boldsymbol{g}(\boldsymbol{x}_k)$ is orthogonal to $\mathcal{N}(\boldsymbol{A}_{\mathcal{I}})$. Since $\Omega_k - \boldsymbol{x}_k \subset \mathcal{N}(\boldsymbol{A}_{\mathcal{I}})$, it follows from Pythagoras that for any $\boldsymbol{z} \in \Omega_k - \boldsymbol{x}_k$, we have

$$\|\boldsymbol{z} + \alpha\boldsymbol{g}(\boldsymbol{x}_k)\|^2 = \|\boldsymbol{z} + \alpha\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)\|^2 + \alpha^2\|\boldsymbol{g}(\boldsymbol{x}_k) - \boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)\|^2.$$

Since $\boldsymbol{z}$ does not appear in the last term, minimizing $\|\boldsymbol{z} + \alpha\boldsymbol{g}(\boldsymbol{x}_k)\|^2$ over $\boldsymbol{z} \in \Omega_k - \boldsymbol{x}_k$ is equivalent to minimizing $\|\boldsymbol{z} + \boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)\|^2$ over $\boldsymbol{z} \in \Omega_k - \boldsymbol{x}_k$. By (4.9), we obtain

$$\boldsymbol{p} = \arg\min\{\|\boldsymbol{z} + \alpha\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)\| : \boldsymbol{z} \in \Omega_k - \boldsymbol{x}_k\}. \tag{4.10}$$

Changing variables from $\boldsymbol{z}$ back to $\boldsymbol{y}$ gives

$$\boldsymbol{p} = \arg\min\{\|\boldsymbol{x}_k - \alpha\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k) - \boldsymbol{y}\| : \boldsymbol{y} \in \Omega_k\} - \boldsymbol{x}_k = \mathcal{P}_{\Omega_k}(\boldsymbol{x}_k - \alpha\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)) - \boldsymbol{x}_k.$$

Comparing this with (4.8) gives (4.6). □

5 Nondegenerate problems

In this section, we focus on the case where the iterates of PASA converge to a nondegenerate stationary point, i.e., a stationary point $\boldsymbol{x}^*$ for which there exists a scalar $\pi > 0$ such that $\lambda_i > \pi$ for all $i \in \mathcal{A}(\boldsymbol{x}^*)$ and $\boldsymbol{\lambda} \in \Lambda(\boldsymbol{x}^*)$.

Theorem 5.1. *If PASA with $\epsilon = 0$ generates an infinite sequence of iterates that converge to a nondegenerate stationary point $\boldsymbol{x}^*$, then within a finite number of iterations, only phase two is executed.*

Proof. By (4.2) and Proposition 4.1, $\boldsymbol{y}(\boldsymbol{x}, 1)$ is close to $\boldsymbol{x}^*$ and $\Lambda(\boldsymbol{x}, 1)$ is close to $\Lambda(\boldsymbol{x}^*)$ when $\boldsymbol{x}$ is close to $\boldsymbol{x}^*$. It follows that for r sufficiently small, we have

$$\lambda_i > 0 \quad \text{for all} \quad i \in \mathcal{A}(\boldsymbol{x}^*), \quad \boldsymbol{\lambda} \in \Lambda(\boldsymbol{x}, 1), \quad \text{and} \quad \boldsymbol{x} \in \mathcal{B}_r(\boldsymbol{x}^*). \tag{5.1}$$

Since $(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})_i < 0$ for all $i \in \mathcal{F}(\boldsymbol{x}^*)$, it also follows from (4.2) that we can take r smaller, if necessary, to ensure that for all $i \in \mathcal{F}(\boldsymbol{x}^*)$ and $\boldsymbol{x} \in \mathcal{B}_r(\boldsymbol{x}^*)$, we have

$$(\boldsymbol{A}\boldsymbol{y}(\boldsymbol{x}, 1) - \boldsymbol{b})_i < 0 \quad \text{and} \quad (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})_i < 0. \tag{5.2}$$

By the last condition in (5.2), we have

$$\mathcal{A}(\boldsymbol{x}) \subset \mathcal{A}(\boldsymbol{x}^*) \quad \text{for all} \quad \boldsymbol{x} \in \mathcal{B}_r(\boldsymbol{x}^*). \tag{5.3}$$

By (5.1) and (5.2),

$$\mathcal{A}(\boldsymbol{y}(\boldsymbol{x}, 1)) = \mathcal{A}(\boldsymbol{x}^*) \quad \text{for all} \quad \boldsymbol{x} \in \mathcal{B}_r(\boldsymbol{x}^*), \tag{5.4}$$

i.e., if $\boldsymbol{x} \in \mathcal{B}_r(\boldsymbol{x}^*)$ and $i \in \mathcal{A}(\boldsymbol{x}^*)$, then by (5.1) and complementary slackness, i lies in $\mathcal{A}(\boldsymbol{y}(\boldsymbol{x}, 1))$, which implies that $\boldsymbol{A}(\boldsymbol{x}^*) \subset \mathcal{A}(\boldsymbol{y}(\boldsymbol{x}, 1))$. Conversely, if $i \in \mathcal{F}(\boldsymbol{x}^*) = \boldsymbol{A}(\boldsymbol{x}^*)^c$, then by (5.2), i lies in $\mathcal{F}(\boldsymbol{y}(\boldsymbol{x}, 1)) = \mathcal{A}(\boldsymbol{y}(\boldsymbol{x}, 1))^c$. Hence, (5.4) holds.

Choose K large enough that $\boldsymbol{x}_k \in \mathcal{B}_r(\boldsymbol{x}^*)$ for all $k \geqslant K$. Since

$$\mathcal{A}(\boldsymbol{x}_k) \subset \mathcal{A}(\boldsymbol{x}^*) = \mathcal{A}(\boldsymbol{y}(\boldsymbol{x}_k, 1))$$

for all $k \geqslant K$ by (5.3) and (5.4), it follows that

$$\mathcal{P}_\Omega(\boldsymbol{x}_k - \boldsymbol{g}(\boldsymbol{x}_k)) = \boldsymbol{y}(\boldsymbol{x}_k, 1) \in \Omega_k \subset \Omega, \tag{5.5}$$

where Ω_k is defined in (4.7). The inclusion (5.5) along with Proposition 4.2 yield

$$\mathcal{P}_\Omega(\boldsymbol{x}_k - \boldsymbol{g}(\boldsymbol{x}_k)) = \mathcal{P}_{\Omega_k}(\boldsymbol{x}_k - \boldsymbol{g}(\boldsymbol{x}_k)) = \mathcal{P}_{\Omega_k}(\boldsymbol{x}_k - \boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)).$$

We subtract $\boldsymbol{x}_k$ from both sides and refer to the definition (2.3) of $\boldsymbol{d}^\alpha$ to obtain

$$\begin{aligned} \boldsymbol{d}^1(\boldsymbol{x}_k) &= \mathcal{P}_{\Omega_k}(\boldsymbol{x}_k - \boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)) - \boldsymbol{x}_k \\ &= \arg\min\{\|\boldsymbol{x}_k - \boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k) - \boldsymbol{y}\|^2 : \boldsymbol{y} \in \Omega_k\} - \boldsymbol{x}_k \\ &= \arg\min\{\|\boldsymbol{z} + \boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)\|^2 : \boldsymbol{z} \in \Omega_k - \boldsymbol{x}_k\}. \end{aligned} \tag{5.6}$$

Recall that at a local minimizer $\bar{\boldsymbol{x}}$ of a smooth function F over the convex set Ω_k, the variational inequality $\nabla F(\bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}}) \geqslant 0$ holds for all $\boldsymbol{x} \in \Omega_k$. We identify F with the objective in (5.6), $\bar{\boldsymbol{x}}$ with $\boldsymbol{d}^1(\boldsymbol{x}_k)$, and $\boldsymbol{x}$ with the point $\boldsymbol{0} \in \Omega_k - \boldsymbol{x}_k$ to obtain the inequality $\boldsymbol{d}^1(\boldsymbol{x}_k)^{\mathrm{T}}(\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k) + \boldsymbol{d}^1(\boldsymbol{x}_k)) \leqslant 0$. Hence,

$$\|\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)\|^2 = \|\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k) + \boldsymbol{d}^1(\boldsymbol{x}_k)\|^2 - 2\boldsymbol{d}^1(\boldsymbol{x}_k)^{\mathrm{T}}(\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k) + \boldsymbol{d}^1(\boldsymbol{x}_k)) + \|\boldsymbol{d}^1(\boldsymbol{x}_k)\|^2 \geqslant \|\boldsymbol{d}^1(\boldsymbol{x}_k)\|^2.$$

By definition, the left-hand side of this inequality is $e(\boldsymbol{x}_k)^2$, while the right-hand side is $E(\boldsymbol{x}_k)^2$. Consequently, $E(\boldsymbol{x}_k) \leqslant e(\boldsymbol{x}_k)$ when $k \geqslant K$. Since $\theta \in (0, 1)$, it follows that phase one immediately branches to phase two, while phase two cannot branch to phase one. This completes the proof. □

6 Degenerate problems

We now focus on a degenerate stationary point $\boldsymbol{x}^*$ where there exist $i \in \mathcal{A}(\boldsymbol{x}^*)$ and $\boldsymbol{\lambda} \in \Lambda(\boldsymbol{x}^*)$ such that $\lambda_i = 0$. We wish to establish a result analogous to Theorem 5.1. To compensate for the degeneracy, it is assumed that the active constraint gradients at $\boldsymbol{x}^*$ are linearly independent, i.e., the rows of $\boldsymbol{A}$ corresponding to indices $i \in \mathcal{A}(\boldsymbol{x}^*)$ are linearly independent, which implies that $\Lambda(\boldsymbol{x}^*, \alpha)$ is a singleton. Under this assumption, Proposition 4.1 yields the following Lipschitz property.

Corollary 6.1. *Suppose $\boldsymbol{x}^*$ is a stationary point for* (1.1) *and the active constraint gradients are linearly independent at $\boldsymbol{x}^*$. If for some $r > 0$, $\boldsymbol{g}$ is Lipschitz continuous in $\mathcal{B}_r(\boldsymbol{x}^*)$ with Lipschitz constant κ and $\boldsymbol{x} \in \mathcal{B}_r(\boldsymbol{x}^*)$ is close enough to $\boldsymbol{x}^*$ that $\mathcal{F}(\boldsymbol{x}^*) \subset \mathcal{F}(\boldsymbol{y}(\boldsymbol{x}, \alpha))$ for some $\alpha \geqslant 0$, then $\Lambda(\boldsymbol{x}, \alpha)$ is a singleton and*

$$\|\Lambda(\boldsymbol{x}, \alpha) - \Lambda(\boldsymbol{x}^*, \alpha)\| \leqslant 2c(1 + \kappa\alpha)\|\boldsymbol{x} - \boldsymbol{x}^*\|,$$

where c is independent of $\boldsymbol{x}$ and depends only on $\boldsymbol{A}$.

Proof. Since $\mathcal{F}(\boldsymbol{x}^*) \subset \mathcal{F}(\boldsymbol{y}(\boldsymbol{x}, \alpha))$, it follows that $\mathcal{A}(\boldsymbol{x}^*) \supset \mathcal{A}(\boldsymbol{y}(\boldsymbol{x}, \alpha))$. Hence, the active constraint gradients are linearly independent at $\boldsymbol{x}^*$ and at $\boldsymbol{y}(\boldsymbol{x}, \alpha)$. This implies that both $\Lambda(\boldsymbol{x}, \alpha)$ and $\Lambda(\boldsymbol{x}^*, \alpha)$ are singletons, and Corollary 6.1 follows from Proposition 4.1. □

To treat degenerate problems, the convergence theory involves one more requirement for the linearly constrained optimizer:

F4. When branching from phase one to phase two, the first iteration in phase two is given by an Armijo line search of the following form: Choose $j \geqslant 0$ as small as possible such that

$$f(\boldsymbol{x}_{k+1}) \leqslant f(\boldsymbol{x}_k) + \delta \nabla f(\boldsymbol{x}_k)(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k), \tag{6.1}$$

where $\boldsymbol{x}_{k+1} = \mathcal{P}_{\Omega_k}(\boldsymbol{x}_k - s_k \boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k))$, $s_k = \alpha\eta^j$, with Ω_k defined in (4.7), $\delta \in (0, 1)$, $\eta \in (0, 1)$, and $\alpha \in (0, \infty)$ (as in the Armijo line search of GPA).

As j grows, η^j tends to zero and $\boldsymbol{x}_{k+1}$ approaches $\boldsymbol{x}_k$. Thus $i \in \mathcal{F}(\boldsymbol{x}_k - \alpha\eta^j \boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k))$ if $i \in \mathcal{F}(\boldsymbol{x}_k)$ and j is large enough. Since $(\boldsymbol{A}\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k))_i = 0$ if $i \in \mathcal{A}(\boldsymbol{x}_k)$, it follows that $\boldsymbol{x}_k - \alpha\eta^j \boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k) \in \Omega_k$ for j sufficiently large, which implies that $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha\eta^j \boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)$; consequently, for j sufficiently large, the Armijo line search inequality (6.1) reduces to the ordinary Armijo line search condition

$$f(\boldsymbol{x}_{k+1}) \leqslant f(\boldsymbol{x}_k) - s_k \delta \nabla f(\boldsymbol{x}_k) \boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k),$$

which holds for s_k sufficiently small. The basic difference between the Armijo line search in F4 and the Armijo line search in GPA is that in F4, the constraints active at $\boldsymbol{x}_k$ remain active at $\boldsymbol{x}_{k+1}$ and F2 holds. With the additional startup procedure F4 for LCO, the global convergence result Theorem 3.2 remains applicable since conditions F1 and F2 are satisfied by the initial iteration in phase two.

Let $\boldsymbol{x}^*$ be a stationary point where the active constraint gradients are linearly independent. For any given $\boldsymbol{x} \in \mathbb{R}^n$, we define

$$\bar{\boldsymbol{x}} = \arg\min_{\boldsymbol{y}} \{\|\boldsymbol{x} - \boldsymbol{y}\| : (\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b})_i = 0 \text{ for all } i \in \mathcal{A}_+(\boldsymbol{x}^*) \cup \mathcal{A}(\boldsymbol{x})\}, \tag{6.2}$$

where $\mathcal{A}_+(\boldsymbol{x}^*) = \{i \in \mathcal{A}(\boldsymbol{x}^*) : \Lambda_i(\boldsymbol{x}^*) > 0\}$. If $\boldsymbol{x}$ is close enough to $\boldsymbol{x}^*$ that $\boldsymbol{A}(\boldsymbol{x}) \subset \boldsymbol{A}(\boldsymbol{x}^*)$, then the feasible set in (6.2) is nonempty since $\boldsymbol{x}^*$ satisfies the constraints; hence, the projection in (6.2) is nonempty when $\boldsymbol{x}$ is sufficiently close to $\boldsymbol{x}^*$.

Lemma 6.2. *Suppose $\boldsymbol{x}^*$ is a stationary point where the active constraint gradients are linearly independent and f is Lipschitz continuously differentiable in a neighborhood of $\boldsymbol{x}^*$. If PASA with $\epsilon = 0$ generates an infinite sequence of iterates $\boldsymbol{x}_k$ converging to $\boldsymbol{x}^*$, then there exists $c \in \mathbb{R}$ such that*

$$\|\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k\| \leqslant c\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \tag{6.3}$$

for all k sufficiently large.

Proof. Choose r in accordance with Corollary 6.1. Similar to what is done in the proof of Theorem 5.1, choose $r > 0$ smaller if necessary to ensure that for all $\boldsymbol{x} \in \mathcal{B}_r(\boldsymbol{x}^*)$, we have

$$\Lambda_i(\boldsymbol{x}, \alpha) > 0 \quad \text{for all} \quad i \in \mathcal{A}_+(\boldsymbol{x}^*), \tag{6.4}$$

and

$$(\boldsymbol{A}\boldsymbol{y}(\boldsymbol{x}, \alpha) - \boldsymbol{b})_j < 0, \quad (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})_j < 0, \quad \text{for all} \quad j \in \mathcal{F}(\boldsymbol{x}^*). \tag{6.5}$$

Choose K large enough that $\boldsymbol{x}_k \in \mathcal{B}_r(\boldsymbol{x}^*)$ for all $k \geqslant K$ and suppose that $\boldsymbol{x}_k$ is any PASA iterate with $k \geqslant K$. If $i \in \mathcal{A}_+(\boldsymbol{x}^*) \cap \mathcal{A}(\boldsymbol{x}_k)$, then by (6.4), $i \in \mathcal{A}(\boldsymbol{y}(\boldsymbol{x}_k, \alpha))$ by complementary slackness. Hence, $i \in \mathcal{A}(\boldsymbol{x})$ for all $\boldsymbol{x}$ on the line segment connecting $\boldsymbol{x}_k$ and $\boldsymbol{y}(\boldsymbol{x}_k, \alpha)$. In particular, $i \in \mathcal{A}(\boldsymbol{x}_{k+1})$ if $\boldsymbol{x}_{k+1}$ is generated by GPA in phase one, while $i \in \mathcal{A}(\boldsymbol{x}_{k+1})$ by F2 if $\boldsymbol{x}_{k+1}$ is generated in phase two. It follows that if constraint $i \in \mathcal{A}_+(\boldsymbol{x}^*)$ becomes active at iterate $\boldsymbol{x}_k$, then $i \in \mathcal{A}(\boldsymbol{x}_l)$ for all $l \geqslant k$. Let $\mathcal{I}$ be the limit of $\mathcal{A}_+(\boldsymbol{x}^*) \cap \mathcal{A}(\boldsymbol{x}_k)$ as k tends to infinity. Choose K larger if necessary to ensure that $\mathcal{I} \subset \mathcal{A}(\boldsymbol{x}_k)$ for all $k \geqslant K$ and suppose that $\boldsymbol{x}_k$ is any iterate of PASA with $k \geqslant K$. If $\mathcal{I} = \mathcal{A}_+(\boldsymbol{x}^*)$, then since $\mathcal{I} \subset \mathcal{A}(\boldsymbol{x}_k)$, it follows that $\mathcal{A}_+(\boldsymbol{x}^*) \cup \mathcal{A}(\boldsymbol{x}_k) = \mathcal{A}(\boldsymbol{x}_k)$, which implies that $\bar{\boldsymbol{x}}_k = \boldsymbol{x}_k$. Thus (6.3) holds trivially since the left-hand side vanishes. Let us focus on the nontrivial case where $\mathcal{I}$ is strictly contained in $\mathcal{A}_+(\boldsymbol{x}^*)$. The analysis is partitioned into three cases.

Case 1. *For k sufficiently large, $\boldsymbol{x}_k$ is generated solely by LCO.* By F3 it follows that for any $\epsilon > 0$, there exists $k \geqslant K$ such that $\|\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)\| = e(\boldsymbol{x}_k) \leqslant \epsilon$. By the first-order optimality conditions for $\boldsymbol{g}^{\mathcal{A}}(\boldsymbol{x}_k)$, there exists $\boldsymbol{\mu}_k \in \mathbb{R}^m$, with $\mu_{ki} = 0$ for all $i \in \mathcal{F}(\boldsymbol{x}_k)$, such that

$$\|\boldsymbol{g}(\boldsymbol{x}_k) + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu}_k\| \leqslant \epsilon. \tag{6.6}$$

The multiplier $\boldsymbol{\mu}_k$ is unique by the independence of the active constraint gradients and the fact that $\mathcal{A}(\boldsymbol{x}_k) \subset \mathcal{A}(\boldsymbol{x}^*)$ by the last condition in (6.5). Similarly, at $\boldsymbol{x}^*$ we have $\boldsymbol{g}(\boldsymbol{x}^*) + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}^* = \boldsymbol{0}$, where $\boldsymbol{\lambda}^* = \Lambda(\boldsymbol{x}^*)$. Combine this with (6.6) to obtain

$$\|\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{\mu}_k - \boldsymbol{\lambda}^*)\| \leqslant \epsilon + \|\boldsymbol{g}(\boldsymbol{x}_k) - \boldsymbol{g}(\boldsymbol{x}^*)\| \leqslant \epsilon + \kappa\|\boldsymbol{x}_k - \boldsymbol{x}^*\|. \tag{6.7}$$

Since $\mathcal{F}(\boldsymbol{x}^*) \subset \mathcal{F}(\boldsymbol{x}_k)$ by the last condition in (6.5), it follows from complementary slackness that $\mu_{ki} = \lambda_i^* = 0$ for all $i \in \mathcal{F}(\boldsymbol{x}^*)$. Since the columns of $\boldsymbol{A}^{\mathrm{T}}$ corresponding to indices in $\mathcal{A}(\boldsymbol{x}^*)$ are linearly independent, there exists a constant c such that

$$\|\boldsymbol{\mu}_k - \boldsymbol{\lambda}^*\| \leqslant c\|\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{\mu}_k - \boldsymbol{\lambda}^*)\|. \tag{6.8}$$

Hence, for ϵ sufficiently small and k sufficiently large, it follows from (6.7) and (6.8) that $\mu_{ki} > 0$ for all $i \in \mathcal{A}_+(\boldsymbol{x}^*)$, which contradicts the assumption that $\mathcal{I}$ is strictly contained in $\mathcal{A}_+(\boldsymbol{x}^*)$. Consequently, Case 1 cannot occur.

Case 2. *PASA makes an infinite number of branches from phase one to phase two and from phase two to phase one.* Let us consider the first iteration of phase two. By Proposition 4.2 and the definition of $\boldsymbol{x}_{k+1}$ in F4, we have $\boldsymbol{x}_{k+1} = \mathcal{P}_{\Omega_k}(\boldsymbol{x}_k - s_k\boldsymbol{g}(\boldsymbol{x}_k))$. The first-order optimality condition for $\boldsymbol{x}_{k+1}$ is that there exists $\boldsymbol{\mu}_k \in \mathbb{R}^m$ such that $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k + s_k\boldsymbol{g}(\boldsymbol{x}_k) + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu}_k = \boldsymbol{0}$, where $\mu_{ki} = 0$ for all $i \in \mathcal{F}(\boldsymbol{x}_{k+1}) \supset \mathcal{F}(\boldsymbol{x}^*)$. Subtracting from this the identity

$$s_k\boldsymbol{g}(\boldsymbol{x}^*) + \boldsymbol{A}^{\mathrm{T}}(s_k\boldsymbol{\lambda}^*) = \boldsymbol{0}$$

yields

$$\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{\mu}_k - s_k\boldsymbol{\lambda}^*) = \boldsymbol{x}_k - \boldsymbol{x}_{k+1} + s_k(\boldsymbol{g}(\boldsymbol{x}^*) - \boldsymbol{g}(\boldsymbol{x}_k)). \tag{6.9}$$

By the Lipschitz continuity of $\boldsymbol{g}$, the bound $s_k \leqslant \alpha$ in F4, and the assumption that the $\boldsymbol{x}_k$ converge to $\boldsymbol{x}^*$, the right-hand side of (6.9) tends to zero as k tends to infinity. Exploiting the independence of the active constraint gradients and the identity $\mu_{ki} = \lambda_i^* = 0$ for all $i \in \mathcal{F}(\boldsymbol{x}^*)$, we deduce from (6.8) and (6.9) that $\|\boldsymbol{\mu}_k - s_k\boldsymbol{\lambda}^*\|$ tends to 0 as k tends to infinity. It follows that for each i, $\mu_{ki} - s_k\lambda_i^*$ tends to zero. If s_k is uniformly bounded away from 0, then $\mu_{ki} > 0$ when $i \in \mathcal{A}_+(\boldsymbol{x}^*)$. By complementary

slackness, $\mathcal{I} = \mathcal{A}_+(\boldsymbol{x}^*)$, which would contradict the assumption that $\mathcal{I}$ is strictly contained in $\mathcal{A}_+(\boldsymbol{x}^*)$. Consequently, Case 2 could not occur.

We will now establish a positive lower bound for s_k in F4 of phase two. If the Armijo stepsize terminates at $j = 0$, then $s_k = \alpha > 0$, and we are done. Next, suppose the stepsize terminates at $j \geqslant 1$. Since j is as small as possible, it follows from Proposition 4.2 and F4 that

$$f(\boldsymbol{x}_k + \boldsymbol{d}_k) - f(\boldsymbol{x}_k) > \delta \boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k, \tag{6.10}$$

where $\boldsymbol{d}_k = \mathcal{P}_{\Omega_k}(\boldsymbol{x}_k - \beta \boldsymbol{g}_k) - \boldsymbol{x}_k$, $\beta := s_k/\eta \leqslant \alpha$. The inequality $\beta \leqslant \alpha$ holds since $j \geqslant 1$. Since $\boldsymbol{x}^* \in \Omega_k \subset \Omega$ by the second condition in (6.5), we have

$$\mathcal{P}_{\Omega_k}(\boldsymbol{x}^* - \beta \boldsymbol{g}(\boldsymbol{x}^*)) = \boldsymbol{x}^*.$$

Since the projection onto a convex set is a nonexpansive operator, we obtain

$$\|(\boldsymbol{x}_k + \boldsymbol{d}_k) - \boldsymbol{x}^*\| = \|\mathcal{P}_{\Omega_k}(\boldsymbol{x}_k - \beta \boldsymbol{g}(\boldsymbol{x}_k)) - \mathcal{P}_{\Omega_k}(\boldsymbol{x}^* - \beta \boldsymbol{g}(\boldsymbol{x}^*))\| \leqslant (1 + \alpha\kappa)\|\boldsymbol{x}_k - \boldsymbol{x}^*\|.$$

The right-hand side of this inequality tends to zero as k tends to infinity. Choose k large enough that $\boldsymbol{x}_k + \boldsymbol{d}_k$ is within the ball centered at $\boldsymbol{x}^*$ where f is Lipschitz continuously differentiable.

Let us expand f in a Taylor series around $\boldsymbol{x}_k$ to obtain

$$\begin{aligned} f(\boldsymbol{x}_k + \boldsymbol{d}_k) - f(\boldsymbol{x}_k) &= \int_0^1 f'(\boldsymbol{x}_k + t\boldsymbol{d}_k)dt = \boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k + \int_0^1 (\nabla f(\boldsymbol{x}_k + t\boldsymbol{d}_k) - \nabla f(\boldsymbol{x}_k))\boldsymbol{d}_k dt \\ &\leqslant \boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k + 0.5\kappa\|\boldsymbol{d}_k\|^2. \end{aligned} \tag{6.11}$$

This inequality combined with (6.10) yields

$$(1 - \delta)\boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k + 0.5\kappa\|\boldsymbol{d}_k\|^2 > 0. \tag{6.12}$$

As in (3.3), but with Ω replaced by Ω_k, we have

$$\boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k \leqslant -\|\boldsymbol{d}_k\|^2/\beta. \tag{6.13}$$

Note that $\boldsymbol{d}_k \neq \boldsymbol{0}$ due to (6.10). Combine (6.12) and (6.13) and replace β by s_k/η to obtain

$$s_k > 2(1 - \delta)\eta/\kappa. \tag{6.14}$$

Hence, if $j \geqslant 1$ in F4, then s_k has the lower bound given in (6.14) for k sufficiently large, while $s_k = \alpha$ if $j = 0$. This completes the proof of Case 2.

Case 3. *For k sufficiently large, $\boldsymbol{x}_k$ is generated solely by GPA.* The Taylor expansion (6.11) can be written as

$$f(\boldsymbol{x}_k + \boldsymbol{d}_k) = f(\boldsymbol{x}_k) + \delta \boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k + (1 - \delta)\boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k + 0.5\kappa\|\boldsymbol{d}_k\|^2, \tag{6.15}$$

where $\boldsymbol{d}_k = \mathcal{P}_{\Omega}(\boldsymbol{x}_k - \alpha \boldsymbol{g}(\boldsymbol{x}_k)) - \boldsymbol{x}_k$ is as defined in GPA. If (6.3) is violated, then for any choice of $c > 0$, there exists $k \geqslant K$ such that

$$\|\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k\| > c\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2. \tag{6.16}$$

By taking c sufficiently large, we will show that

$$(1 - \delta)\boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k + 0.5\kappa\|\boldsymbol{d}_k\|^2 \leqslant 0. \tag{6.17}$$

In this case, (6.15) implies that $s_k = 1$ is accepted in GPA and

$$\boldsymbol{x}_{k+1} = \mathcal{P}_{\Omega}(\boldsymbol{x}_k - \alpha \boldsymbol{g}(\boldsymbol{x}_k)).$$

By Corollary 6.1, $\|\Lambda(\boldsymbol{x}_k, \alpha) - \Lambda(\boldsymbol{x}^*, \alpha)\| = \|\Lambda(\boldsymbol{x}_k, \alpha) - \alpha\boldsymbol{\lambda}^*\|$ tends to 0 as k tends to infinity. This implies that $\Lambda_i(\boldsymbol{x}_k, \alpha) > 0$ when $\lambda_i^* > 0$, which contradicts the assumption that $\mathcal{I}$ is strictly contained in $\mathcal{A}_+(\boldsymbol{x}^*)$. Hence, (6.3) cannot be violated.

To establish (6.17), first observe that

$$\|\boldsymbol{d}_k\| = \|\boldsymbol{y}(\boldsymbol{x}_k,\alpha) - \boldsymbol{x}_k\| \leqslant \|\boldsymbol{y}(\boldsymbol{x}_k,\alpha) - \boldsymbol{x}^*\| + \|\boldsymbol{x}^* - \boldsymbol{x}_k\| \leqslant (2+\alpha\kappa)\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \tag{6.18}$$

by (4.2). By the first-order optimality condition (4.4) for $\boldsymbol{y}(\boldsymbol{x}_k,\alpha)$, it follows that

$$\boldsymbol{d}_k = -(\alpha\boldsymbol{g}(\boldsymbol{x}_k) + \boldsymbol{A}^{\mathrm{T}}\Lambda(\boldsymbol{x}_k,\alpha)).$$

The dot product of this equation with $\boldsymbol{d}_k$ gives

$$(\alpha\boldsymbol{g}_k + \boldsymbol{A}^{\mathrm{T}}\Lambda(\boldsymbol{x}_k,\alpha))^{\mathrm{T}}\boldsymbol{d}_k = -\|\boldsymbol{d}_k\|^2 \leqslant 0. \tag{6.19}$$

Again, by the definition of $\boldsymbol{d}_k$ and by complementary slackness, we have

$$\Lambda(\boldsymbol{x}_k,\alpha)^{\mathrm{T}}\boldsymbol{A}\boldsymbol{d}_k = \Lambda(\boldsymbol{x}_k,\alpha)^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{y}(\boldsymbol{x}_k,\alpha) - \boldsymbol{x}_k) = \Lambda(\boldsymbol{x}_k,\alpha)^{\mathrm{T}}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k). \tag{6.20}$$

By Corollary 6.1, it follows that for K sufficiently large and for any $k \geqslant K$,

$$\Lambda_i(\boldsymbol{x}_k,\alpha) \geqslant 0.5\Lambda_i(\boldsymbol{x}^*,\alpha) \quad \text{for all} \quad i \in \mathcal{A}_+(\boldsymbol{x}^*).$$

Hence, for any $i \in \mathcal{A}_+(\boldsymbol{x}^*)$ and $k \geqslant K$, (6.20) gives

$$\Lambda(\boldsymbol{x}_k,\alpha)^{\mathrm{T}}\boldsymbol{A}\boldsymbol{d}_k \geqslant 0.5\Lambda_i(\boldsymbol{x}^*,\alpha)(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)_i = 0.5\alpha\Lambda_i(\boldsymbol{x}^*)(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)_i \tag{6.21}$$

since $\Lambda(\boldsymbol{x}_k,\alpha) \geqslant \boldsymbol{0}$, $\boldsymbol{A}\boldsymbol{x}_k \leqslant \boldsymbol{b}$, and each term in the inner product $\Lambda(\boldsymbol{x}_k,\alpha)^{\mathrm{T}}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)$ is nonnegative. Combine (6.19)–(6.21) to obtain

$$\boldsymbol{g}_k^{\mathrm{T}}\boldsymbol{d}_k \leqslant -0.5\Lambda_i(\boldsymbol{x}^*)(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)_i \tag{6.22}$$

for any $i \in \mathcal{A}_+(\boldsymbol{x}^*)$ and $k \geqslant K$. The distance $\|\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k\|$ between $\boldsymbol{x}_k$ and its projection $\bar{\boldsymbol{x}}_k$ in (6.2) is bounded by a constant times the maximum violation of the constraint $(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)_i = 0$ for $i \in \mathcal{A}_+(\boldsymbol{x}^*) \cup \mathcal{A}(\boldsymbol{x}_k)$, i.e., there exists a constant $\bar{c}$ such that

$$\|\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k\| \leqslant \bar{c}\max\{(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)_i : i \in \mathcal{A}_+(\boldsymbol{x}^*) \cup \mathcal{A}(\boldsymbol{x}_k)\}. \tag{6.23}$$

Since $(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)_i = 0$ for all $i \in \mathcal{A}(\boldsymbol{x}_k)$, it follows that the maximum constraint violation in (6.23) is achieved for some $i \in \mathcal{A}_+(\boldsymbol{x}^*)$ (otherwise, $\bar{\boldsymbol{x}}_k = \boldsymbol{x}_k$ and (6.16) is violated). Consequently, if the index $i \in \mathcal{A}_+(\boldsymbol{x}^*)$ in (6.22) is chosen to make $(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)_i$ as large as possible, then

$$\boldsymbol{g}_k^{\mathrm{T}}\boldsymbol{d}_k \leqslant -d\|\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k\|,$$

where $d = 0.5\bar{c}\min\{\Lambda_i(\boldsymbol{x}^*) : i \in \mathcal{A}_+(\boldsymbol{x}^*)\}$. If (6.16) holds, then by (6.18), we have

$$\boldsymbol{g}_k^{\mathrm{T}}\boldsymbol{d}_k \leqslant -cd\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \leqslant -cd\|\boldsymbol{d}_k\|^2/(2+\alpha\kappa).$$

Hence, the expression (6.17) has the upper bound

$$(1-\delta)\boldsymbol{g}_k^{\mathrm{T}}\boldsymbol{d}_k + 0.5\kappa\|\boldsymbol{d}_k\|^2 \leqslant \left(\frac{cd(\delta-1)}{2+\alpha\kappa} + 0.5\kappa\right)\|\boldsymbol{d}_k\|^2.$$

Since $\delta < 1$, this is nonpositive when c is sufficiently large. This completes the proof of (6.17). □

Note that there is a fundamental difference between the prototype GPA used in this paper and the versions of the gradient projection algorithm based on a piecewise projected gradient such as those in [4–6]. In GPA there is a single projection followed by a backtrack towards the starting point. Consequently, we are unable to show that the active constraints are identified in a finite number of iterations, unlike the piecewise projection schemes, where the active constraints can be identified in a finite number of iterations, but at the expense additional projections when the stepsize increases. In Lemma 6.2, we show that even though we do not identify the active constraints, the violation of the constraints $(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})_i = 0$ for $i \in \mathcal{A}_+(\boldsymbol{x}^*)$ by iterate $\boldsymbol{x}_k$ is on the order of the error in $\boldsymbol{x}_k$ squared.

When $\boldsymbol{x}^*$ is fully determined by the active constraints for which the strict complementarity holds, convergence is achieved in a finite number of iterations as we now show.

Corollary 6.3. *Suppose $\boldsymbol{x}^*$ is a stationary point where the active constraint gradients are linearly independent and f is Lipschitz continuously differentiable in a neighborhood of $\boldsymbol{x}^*$. If the PASA iterates $\boldsymbol{x}_k$ converge to $\boldsymbol{x}^*$ and $|\mathcal{A}_+(\boldsymbol{x}^*)| = n$, then $\boldsymbol{x}_k = \boldsymbol{x}^*$ after a finite number of iterations.*

Proof. Choose k large enough that $\mathcal{A}(\boldsymbol{x}_k) \subset \mathcal{A}(\boldsymbol{x}^*)$ and $\bar{\boldsymbol{x}}_k$ is nonempty. Since $|\mathcal{A}_+(\boldsymbol{x}^*)| = n$ and the active constraint gradients are linearly independent, we have $\bar{\boldsymbol{x}}_k = \boldsymbol{x}^*$. By Lemma 6.2, we must have $\boldsymbol{x}_k = \boldsymbol{x}^*$ whenever $\|\boldsymbol{x}_k - \boldsymbol{x}^*\| < 1/c$. □

To complete the analysis of PASA in the degenerate case and show that PASA ultimately performs only iterations in phase two, we also need to assume that the strong second-order sufficient optimality condition holds. Recall that a stationary point $\boldsymbol{x}^*$ of (1.1) satisfies the strong second-order sufficient optimality condition if there exists $\sigma > 0$ such that

$$\boldsymbol{d}^{\mathrm{T}}\nabla^2 f(\boldsymbol{x}^*)\boldsymbol{d} \geqslant \sigma\|\boldsymbol{d}\|^2 \quad \text{whenever} \quad (\boldsymbol{A}\boldsymbol{d})_i = 0 \quad \text{for all} \quad i \in \mathcal{A}_+(\boldsymbol{x}^*). \tag{6.24}$$

First, we observe that under this assumption, the distance from $\boldsymbol{x}_k$ to $\boldsymbol{x}^*$ is bounded in terms of $E(\boldsymbol{x}_k)$.

Lemma 6.4. *If f is twice continuously differentiable in a neighborhood of a local minimizer $\boldsymbol{x}^*$ for* (1.1) *where the active constraint gradients are linearly independent and the strong second-order sufficient optimality condition holds, then for some $\rho > 0$ and for all $\boldsymbol{x} \in \mathcal{B}_\rho(\boldsymbol{x}^*)$, we have*

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leqslant \left[\sqrt{1 + \left(\frac{2(1+\kappa)(3+\kappa)}{\sigma}\right)^2}\,\right] E(\boldsymbol{x}), \tag{6.25}$$

where κ is a Lipschitz constant for ∇f on $\mathcal{B}_\rho(\boldsymbol{x}^*)$.

Proof. By the continuity of the second derivative of f, it follows from (6.24) that for $\rho > 0$ sufficiently small,

$$(\boldsymbol{x} - \boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}^*)) = (\boldsymbol{x} - \boldsymbol{x}^*)^{\mathrm{T}} \int_0^1 \nabla^2 f(\boldsymbol{x}^* + t(\boldsymbol{x} - \boldsymbol{x}^*))dt(\boldsymbol{x} - \boldsymbol{x}^*) \geqslant 0.5\sigma\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \tag{6.26}$$

for all $\boldsymbol{x} \in \mathcal{B}_\rho(\boldsymbol{x}^*) \cap \mathcal{S}_+$, where

$$\mathcal{S}_+ = \{\boldsymbol{x} \in \mathbb{R}^n : (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})_i = 0 \text{ for all } i \in \mathcal{A}_+(\boldsymbol{x}^*)\}.$$

Given $\boldsymbol{x} \in \mathcal{B}_\rho(\boldsymbol{x}^*)$, define $\hat{\boldsymbol{x}} = \mathcal{P}_{\mathcal{S}_+}(\boldsymbol{x})$. Since $\mathcal{P}_{\Omega\cap\mathcal{S}_+}(\boldsymbol{x} - \boldsymbol{g}(\boldsymbol{x})) \in \mathcal{S}_+$, it follows that

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\| = \|\mathcal{P}_{\mathcal{S}_+}(\boldsymbol{x}) - \boldsymbol{x}\| \leqslant \|\mathcal{P}_{\Omega\cap\mathcal{S}_+}(\boldsymbol{x} - \boldsymbol{g}(\boldsymbol{x})) - \boldsymbol{x}\|. \tag{6.27}$$

Since $\Lambda_i(\boldsymbol{x}^*) > 0$ for all $i \in \mathcal{A}_+(\boldsymbol{x}^*)$, it follows from Corollary 6.1 and complementary slackness that ρ can be chosen smaller if necessary to ensure that

$$(\boldsymbol{A}\boldsymbol{y}(\boldsymbol{x}, 1) - \boldsymbol{b})_i = 0 \quad \text{for all} \quad i \in \mathcal{A}_+(\boldsymbol{x}^*), \tag{6.28}$$

which implies that $\boldsymbol{y}(\boldsymbol{x}, 1) \in \mathcal{S}_+$. Since $\boldsymbol{y}(\boldsymbol{x}, 1) = \mathcal{P}_\Omega(\boldsymbol{x} - \boldsymbol{g}(\boldsymbol{x}))$ and $\boldsymbol{y}(\boldsymbol{x}, 1) \in \mathcal{S}_+$, we also have $\mathcal{P}_\Omega(\boldsymbol{x} - \boldsymbol{g}(\boldsymbol{x})) = \mathcal{P}_{\Omega\cap\mathcal{S}_+}(\boldsymbol{x} - \boldsymbol{g}(\boldsymbol{x}))$. With this substitution in (6.27), we obtain

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\| \leqslant \|\mathcal{P}_\Omega(\boldsymbol{x} - \boldsymbol{g}(\boldsymbol{x})) - \boldsymbol{x}\| = \|\boldsymbol{y}(\boldsymbol{x}, 1) - \boldsymbol{x}\| = \|\boldsymbol{d}^1(\boldsymbol{x})\| = E(\boldsymbol{x}). \tag{6.29}$$

By the Lipschitz continuity of $\boldsymbol{g}$, (6.29), and (4.3), it follows that

$$\|\boldsymbol{d}^1(\hat{\boldsymbol{x}})\| \leqslant \|\boldsymbol{d}^1(\boldsymbol{x})\| + \|\boldsymbol{d}^1(\hat{\boldsymbol{x}}) - \boldsymbol{d}^1(\boldsymbol{x})\| \leqslant \|\boldsymbol{d}^1(\boldsymbol{x})\| + (2+\kappa)\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| \leqslant (3+\kappa)\|\boldsymbol{d}^1(\boldsymbol{x})\| \tag{6.30}$$

for all $\boldsymbol{x} \in \mathcal{B}_\rho(\boldsymbol{x}^*)$. Since $\hat{\boldsymbol{x}} = \mathcal{P}_{\mathcal{S}_+}(\boldsymbol{x})$, the difference $\hat{\boldsymbol{x}} - \boldsymbol{x}$ is orthogonal to $\mathcal{N}(\boldsymbol{A}_\mathcal{I})$ when $\mathcal{I} = \mathcal{A}_+(\boldsymbol{x}^*)$. Since $\hat{\boldsymbol{x}} - \boldsymbol{x}^* \in \mathcal{N}(\boldsymbol{A}_\mathcal{I})$, it follows from Pythagoras that

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 + \|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\|^2 = \|\boldsymbol{x} - \boldsymbol{x}^*\|^2. \tag{6.31}$$

Consequently, $\hat{\boldsymbol{x}} \in \mathcal{B}_\rho(\boldsymbol{x}^*)$ for all $\boldsymbol{x} \in \mathcal{B}_\rho(\boldsymbol{x}^*)$, and

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| = \sqrt{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 + \|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\|^2}. \tag{6.32}$$

By [27, p. 8], (6.26), and (6.30), we have

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\| \leqslant \left(\frac{1+\kappa}{0.5\sigma}\right)\|\boldsymbol{d}^1(\hat{\boldsymbol{x}})\| \leqslant \left(\frac{(1+\kappa)(3+\kappa)}{0.5\sigma}\right)\|\boldsymbol{d}^1(\boldsymbol{x})\|. \tag{6.33}$$

Insert (6.29) and (6.33) in (6.32) to complete the proof. □

We now examine the asymptotic behavior of the undecided index set $\mathcal{U}$.

Lemma 6.5. *If f is twice continuously differentiable in a neighborhood of a local minimizer $\boldsymbol{x}^*$ for* (1.1)*, where the active constraint gradients are linearly independent and the strong second-order sufficient optimality condition holds, and if PASA generates an infinite sequence of iterates converging to $\boldsymbol{x}^*$, then $\mathcal{U}(\boldsymbol{x}_k)$ is empty for k sufficiently large.*

Proof. If $E(\boldsymbol{x}_k) = 0$ for some k, then PASA terminates and the lemma holds trivially. Hence, assume that $E(\boldsymbol{x}_k) \neq 0$ for all k. To show $\mathcal{U}(\boldsymbol{x}_k)$ is empty for some k, we must show that

$$\text{either} \quad \text{(a)} \quad \boldsymbol{\lambda}_i(\boldsymbol{x}_k)/E(\boldsymbol{x}_k)^\gamma < 1 \quad \text{or} \quad \text{(b)} \quad (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})_i/E(\boldsymbol{x})^\beta < 1$$

for each i. If $i \in \mathcal{A}_+(\boldsymbol{x}^*)$, then $[\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{x}}_k]_i = 0$, and by Lemma 6.2,

$$[\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k]_i = [(\boldsymbol{A}(\bar{\boldsymbol{x}}_k - \boldsymbol{x}_k)]_i \leqslant \|\boldsymbol{A}\|\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}_k\| \leqslant c\|\boldsymbol{A}\|\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2.$$

By Lemma 6.4, there exists a constant d such that $\|\boldsymbol{x} - \boldsymbol{x}^*\| \leqslant dE(\boldsymbol{x})$ for $\boldsymbol{x}$ near $\boldsymbol{x}^*$. Hence, for all $i \in \mathcal{A}_+(\boldsymbol{x}^*)$ and k sufficiently large, we have

$$[\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k]_i \leqslant cd^2\|\boldsymbol{A}\|E(\boldsymbol{x}_k)^2 = cd^2\|\boldsymbol{A}\|E(\boldsymbol{x}_k)^{2-\beta}E(\boldsymbol{x}_k)^\beta.$$

Since $\beta \in (1,2)$, $E(\boldsymbol{x}_k)^{2-\beta}$ tends to zero as k tends to infinity, and (b) holds when k is large enough that $cd^2\|\boldsymbol{A}\|E(\boldsymbol{x}_k)^{2-\beta} < 1$.

If $i \in \mathcal{A}_+(\boldsymbol{x}^*)^c$, then $\lambda_i(\boldsymbol{x}^*) = 0$. By Corollary 6.1, there exist $c \in \mathbb{R}$ such that

$$\lambda_i(\boldsymbol{x}_k) = \lambda_i(\boldsymbol{x}_k) - \lambda_i(\boldsymbol{x}^*) \leqslant c\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leqslant cdE(\boldsymbol{x}_k) = cdE(\boldsymbol{x}_k)^{1-\gamma}E(\boldsymbol{x}_k)^\gamma.$$

Since $\gamma \in (0,1)$, $E(\boldsymbol{x}_k)^{1-\gamma}$ tends to zero as k tends to infinity, and (a) holds when k is large enough that $cdE(\boldsymbol{x}_k)^{1-\gamma} < 1$. In summary, for k sufficiently large, (a) holds when $i \in \mathcal{A}_+(\boldsymbol{x}^*)^c$ and (b) holds when $i \in \mathcal{A}_+(\boldsymbol{x}^*)$. This implies that $\mathcal{U}(\boldsymbol{x}_k)$ is empty for k sufficiently large. □

As shown in the proof of Lemma 6.5, (b) holds for $i \in \mathcal{A}_+(\boldsymbol{x}^*)$ and k sufficiently large. This implies that the constraint violation $(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})_i$ tends to zero faster than the error $E(\boldsymbol{x}_k)$. The following result, along with Lemma 6.5, essentially implies that PASA eventually performs only phase two.

Lemma 6.6. *If f is twice continuously differentiable in a neighborhood of a local minimizer $\boldsymbol{x}^*$ for* (1.1)*, where the active constraint gradients are linearly independent and the strong second-order sufficient optimality condition holds, and if PASA generates an infinite sequence of iterates converging to $\boldsymbol{x}^*$, then there exists $\theta^* > 0$ such that*

$$e(\boldsymbol{x}_k) \geqslant \theta^* E(\boldsymbol{x}_k) \tag{6.34}$$

for k sufficiently large.

Proof. Let $\mathcal{I} := \mathcal{A}_+(\boldsymbol{x}^*) \cup \mathcal{A}(\boldsymbol{x}_k)$. The projection $\bar{\boldsymbol{x}}_k$ has the property that the difference $\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k$ is orthogonal to $\mathcal{N}(\boldsymbol{A}_\mathcal{I})$. Choose k large enough that $\mathcal{A}(\boldsymbol{x}_k) \subset \mathcal{A}(\boldsymbol{x}^*)$. It follows that $\bar{\boldsymbol{x}}_k - \boldsymbol{x}^* \in \mathcal{N}(\boldsymbol{A}_\mathcal{I})$. Hence, by Pythagoras, we have

$$\|\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k\|^2 + \|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|^2 = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2.$$

Consequently,

$$\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\| \leqslant \|\boldsymbol{x}_k - \boldsymbol{x}^*\|, \tag{6.35}$$

and $\bar{\boldsymbol{x}}_k$ approaches $\boldsymbol{x}^*$ as k tends to infinity. Choose $\rho > 0$ small enough that f is twice continuously differentiable in $\mathcal{B}_\rho(\boldsymbol{x}^*)$, and let κ be the Lipschitz constant for ∇f in $\mathcal{B}_\rho(\boldsymbol{x}^*)$. Choose k large enough that $\boldsymbol{x}_k \in \mathcal{B}_\rho(\boldsymbol{x}^*)$. By (6.35) $\bar{\boldsymbol{x}}_k \in \mathcal{B}_\rho(\boldsymbol{x}^*)$. Since $\boldsymbol{d}^1(\boldsymbol{x}^*) = \boldsymbol{0}$, it follows from (4.3) that

$$\begin{aligned}\|\boldsymbol{d}^1(\boldsymbol{x}_k)\| &\leqslant \|\boldsymbol{d}^1(\boldsymbol{x}_k) - \boldsymbol{d}^1(\bar{\boldsymbol{x}}_k)\| + \|\boldsymbol{d}^1(\bar{\boldsymbol{x}}_k) - \boldsymbol{d}^1(\boldsymbol{x}^*)\| \\ &\leqslant (2+\kappa)(\|\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k\| + \|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|).\end{aligned} \tag{6.36}$$

Lemma 6.2 gives

$$\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}_k\| \leqslant c\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \leqslant c\|\boldsymbol{x}_k - \boldsymbol{x}^*\|(\|\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k\| + \|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|). \tag{6.37}$$

Since $\boldsymbol{x}_k$ converges to $\boldsymbol{x}^*$, it follows from (6.37) that for any $\epsilon > 0$,

$$\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}_k\| \leqslant \epsilon\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\| \tag{6.38}$$

when k is sufficiently large. Combine (6.36) and (6.38) to obtain

$$\|\boldsymbol{d}^1(\boldsymbol{x}_k)\| \leqslant c\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\| \tag{6.39}$$

for some constant c and any k sufficiently large.

Choose $\rho > 0$ small enough that (6.26) holds for all $\boldsymbol{x} \in \mathcal{B}_\rho(\boldsymbol{x}^*)$, and choose k large enough that $\bar{\boldsymbol{x}}_k \in \mathcal{B}_\rho(\boldsymbol{x}^*)$. The bound (6.26) yields

$$0.5\sigma\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|^2 \leqslant (\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{g}(\bar{\boldsymbol{x}}_k) - \boldsymbol{g}(\boldsymbol{x}^*)). \tag{6.40}$$

By the first-order optimality conditions for a local minimizer $\boldsymbol{x}^*$ of (1.1), there exists a multiplier $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ such that

$$\boldsymbol{g}(\boldsymbol{x}^*) + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}^* = 0, \tag{6.41}$$

where $(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{\lambda}^* = 0$ and $\boldsymbol{\lambda}^* \geqslant \boldsymbol{0}$. Observe that $\lambda_i^*[\boldsymbol{A}(\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*)]_i = 0$ for each i since $[\boldsymbol{A}(\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*)]_i = 0$ when $i \in \mathcal{A}_+(\boldsymbol{x}^*)$, while $\lambda_i^* = 0$ when $i \in \mathcal{A}_+(\boldsymbol{x}^*)^c$. Hence, we have

$$[\boldsymbol{A}(\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*)]^{\mathrm{T}}\boldsymbol{\lambda}^* = 0.$$

We utilize this identity to obtain

$$(\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{x}^*) = (\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{g}(\boldsymbol{x}^*) + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}^*) = \boldsymbol{0} \tag{6.42}$$

by the first equality in (6.41).

The first-order optimality conditions for the minimizer $\boldsymbol{g}^{\mathcal{I}}(\bar{\boldsymbol{x}}_k)$ in (2.1) imply the existence of $\boldsymbol{\lambda}_{\mathcal{I}}$ such that

$$\boldsymbol{g}^{\mathcal{I}}(\bar{\boldsymbol{x}}_k) - \boldsymbol{g}(\bar{\boldsymbol{x}}_k) + \boldsymbol{A}_{\mathcal{I}}^{\mathrm{T}}\boldsymbol{\lambda}_{\mathcal{I}} = \boldsymbol{0}, \tag{6.43}$$

where $\boldsymbol{A}_{\mathcal{I}}\bar{\boldsymbol{x}}_k = \boldsymbol{b}_{\mathcal{I}}$. Since $\mathcal{A}(\boldsymbol{x}_k) \subset \mathcal{A}(\boldsymbol{x}^*)$, we have $\boldsymbol{A}_{\mathcal{I}}(\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*) = \boldsymbol{0}$, $[\boldsymbol{A}_{\mathcal{I}}(\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*)]^{\mathrm{T}}\boldsymbol{\lambda}_{\mathcal{I}} = \boldsymbol{0}$, and

$$(\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{g}(\bar{\boldsymbol{x}}_k) = (\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{g}(\bar{\boldsymbol{x}}_k) - \boldsymbol{A}_{\mathcal{I}}^{\mathrm{T}}\boldsymbol{\lambda}_{\mathcal{I}}) = (\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{g}^{\mathcal{I}}(\bar{\boldsymbol{x}}_k) \tag{6.44}$$

by (6.43). Combine (6.40), (6.42), and (6.44) to obtain

$$0.5\sigma\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\| \leqslant \|\boldsymbol{g}^{\mathcal{I}}(\bar{\boldsymbol{x}}_k)\|. \tag{6.45}$$

If $\mathcal{J}$ denotes $\mathcal{A}(\boldsymbol{x}_k)$, then $\mathcal{J} \subset \mathcal{I} = \mathcal{A}(\boldsymbol{x}_k) \cup \mathcal{A}_+(\boldsymbol{x}_k)$. Hence, $\mathcal{N}(\boldsymbol{A}_{\mathcal{I}}) \subset \mathcal{N}(\boldsymbol{A}_{\mathcal{J}})$. It follows that

$$\|\boldsymbol{g}^{\mathcal{I}}(\boldsymbol{x}_k)\| \leqslant \|\boldsymbol{g}^{\mathcal{J}}(\boldsymbol{x}_k)\| = e(\boldsymbol{x}_k). \tag{6.46}$$

Since the projection on a convex set is nonexpansive,

$$\|\boldsymbol{g}^{\mathcal{I}}(\bar{\boldsymbol{x}}_k) - \boldsymbol{g}^{\mathcal{I}}(\boldsymbol{x}_k)\| \leqslant \|\boldsymbol{g}(\bar{\boldsymbol{x}}_k) - \boldsymbol{g}(\boldsymbol{x}_k)\| \leqslant \kappa\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}_k\|. \tag{6.47}$$

Combine (6.38), (6.46), and (6.47) to get

$$\begin{aligned}\|\boldsymbol{g}^{\mathcal{I}}(\bar{\boldsymbol{x}}_k)\| &\leqslant \|\boldsymbol{g}^{\mathcal{I}}(\bar{\boldsymbol{x}}_k) - \boldsymbol{g}^{\mathcal{I}}(\boldsymbol{x}_k)\| + \|\boldsymbol{g}^{\mathcal{I}}(\boldsymbol{x}_k)\| \\ &\leqslant e(\boldsymbol{x}_k) + \kappa\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}_k\| \leqslant e(\boldsymbol{x}_k) + \epsilon\kappa\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\|.\end{aligned}$$

Consequently, by (6.45) we have $0.4\sigma\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}^*\| \leqslant e(\boldsymbol{x}_k)$ for ϵ sufficiently small and k sufficiently large. Finally, (6.39) completes the proof. □

By the analysis of Section 5, (6.34) holds with $\theta^* = 1$ for a nondegenerate problem; neither the strong second-order sufficient optimality condition nor independence of the active constraint gradients are needed in this case.

We now show that within a finite number of iterations, PASA will perform only LCO.

Theorem 6.7. *If PASA with $\epsilon = 0$ generates an infinite sequence of iterates converging to a local minimizer $\boldsymbol{x}^*$ of* (1.1)*, where the active constraint gradients are linearly independent and the strong second-order sufficient optimality condition holds, and if f is twice continuously differentiable near $\boldsymbol{x}^*$, then within a finite number of iterations, only phase two is executed.*

Proof. By Lemma 6.5, the undecided index set $\mathcal{U}(\boldsymbol{x}_k)$ is empty for k sufficiently large, and by Lemma 6.6, there exists $\theta^* > 0$ such that $e(\boldsymbol{x}_k) \geqslant \theta^* E(\boldsymbol{x}_k)$. If k is large enough that $\mathcal{U}(\boldsymbol{x}_k)$ is empty, then in phase one, θ will be reduced until $\theta \leqslant \theta^*$. Once this holds, phase one branches to phase two and phase two cannot branch to phase one. □

Similar to [27, Theorem 4.2], when f is a strongly convex quadratic and LCO is based on a projected conjugate gradient method, Theorem 6.7 implies that when the active constraint gradients are linearly independent, PASA converges to the optimal solution in a finite number of iterations.

7 Conclusions

A new active set algorithm PASA was developed for solving polyhedral constrained nonlinear optimization problems. Phase one of the algorithm is the gradient projection algorithm, while phase two is any algorithm for linearly constrained optimization (LCO) which monotonically improves the value of the objective function, which never frees an active constraint, and which has the property that the projected gradients tend to zero, at least along a subsequence of the iterates. Simple rules were given in Algorithm 2 for branching between the two phases. Global convergence to a stationary point was established, while asymptotically, within a finite number of iterations, only phase two is performed. For nondegenerate problems, this result follows almost immediately, while for degenerate problems, the analysis required linear independence of the active constraint gradients, the strong second-order sufficient optimality conditions, and a special startup procedure for LCO. The numerical implementation and performance of PASA for general polyhedral constrained problems will be studied in a separate paper. Numerical performance for bound constrained optimization problems is studied in [27].

Acknowledgements This work was supported by the National Science Foundation of USA (Grant Nos. 1522629 and 1522654), the Office of Naval Research of USA (Grant Nos. N00014-11-1-0068 and N00014-15-1-2048), the Air Force Research Laboratory of USA (Contract No. FA8651-08-D-0108/0054), and National Natural Science Foundation of China (Grant No. 11571178).

References

1 Bertsekas D P. On the Goldstein-Levitin-Polyak gradient projection method. IEEE Trans Automat Control, 1976, 21: 174–184

2 Bertsekas D P. Projected Newton methods for optimization problems with simple constraints. SIAM J Control Optim, 1982, 20: 221–246
3 Branch M, Coleman T, Li Y. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. SIAM J Sci Comput, 1999, 21: 1–23
4 Burke J V, Moré J J. On the identification of active constraints. SIAM J Numer Anal, 1988, 25: 1197–1211
5 Burke J V, Moré J J. Exposing constraints. SIAM J Optim, 1994, 25: 573–595
6 Burke J V, Moré J J, Toraldo G. Convergence properties of trust region methods for linear and convex constraints. Math Program, 1990, 47: 305–336
7 Calamai P, Moré J. Projected gradient for linearly constrained problems. Math Program, 1987, 39: 93–116
8 Coleman T F, Li Y. On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds. Math Program, 1994, 67: 189–224
9 Coleman T F, Li Y. An interior trust region approach for nonlinear minimization subject to bounds. SIAM J Optim, 1996, 6: 418–445
10 Coleman T F, Li Y. A trust region and affine scaling interior point method for nonconvex minimization with linear inequality constraints. Math Program, 2000, 88: 1–31
11 Conn A R, Gould N I M, Toint P L. Global convergence of a class of trust region algorithms for optimization with simple bounds. SIAM J Numer Anal, 1988, 25: 433–460
12 Conn A R, Gould N I M, Toint P L. Trust-Region Methods. Philadelphia: SIAM, 2000
13 Dembo R S, Tulowitzki U. On the minimization of quadratic functions subject to box constraints. Working paper. New Haven: Yale University, 1983
14 Dennis J E, Heinkenschloss M, Vicente L N. Trust-region interior-point algorithms for a class of nonlinear programming problems. SIAM J Control Optim, 1998, 36: 1750–1794
15 Dostál Z. Box constrained quadratic programming with proportioning and projections. SIAM J Optim, 1997, 7: 871–887
16 Dostál Z. A proportioning based algorithm for bound constrained quadratic programming with the rate of convergence. Numer Algorithms, 2003, 34: 293–302
17 Dostál Z, Friedlander A, Santos S A. Augmented Lagrangians with adaptive precision control for quadratic programming with simple bounds and equality constraints. SIAM J Optim, 2003, 13: 1120–1140
18 Facchinei F, Júdice J, Soares J. An active set Newton's algorithm for large-scale nonlinear programs with box constraints. SIAM J Optim, 1998, 8: 158–186
19 Facchinei F, Lucidi S, Palagi L. A truncated Newton algorithm for large-scale box constrained optimization. SIAM J Optim, 2002, 4: 1100–1125
20 Forsgren A, Gill P E, Wong E. Active-set methods for convex quadratic programming. ArXiv:1503.08349, 2015
21 Friedlander A, Martínez J M, Santos S A. A new trust region algorithm for bound constrained minimization. Appl Math Optim, 1994, 30: 235–266
22 Friedlander M P, Leyffer S. Global and finite termination of a two-phase augmented Lagrangian filter method for general quadratic programs. SIAM J Sci Comput, 2008, 30: 1706–1729
23 Gill P E, Wong E. Methods for convex and general quadratic programming. Math Prog Comp, 2014, 7: 71–112
24 Goldberg N, Leyffer S. An active-set method for second-order conic-constrained quadratic programming. SIAM J Optim, 2015, 25: 1455–1477
25 Goldstein A A. Convex programming in Hilbert space. Bull Amer Math Soc, 1964, 70: 709–710
26 Hager W W, Phan D T, Zhang H. Gradient-based methods for sparse recovery. SIAM J Imaging Sci, 2011, 4: 146–165
27 Hager W W, Zhang H. A new active set algorithm for box constrained optimization. SIAM J Optim, 2006, 17: 526–557
28 Hager W W, Zhang H. An affine scaling method for optimization problems with polyhedral constraints. Comput Optim Appl, 2014, 59: 163–183
29 Hager W W, Zhang H. Projection on a polyhedron that exploits sparsity. SIAM J Optim, 2015, in press
30 Heinkenschloss M, Ulbrich M, Ulbrich S. Superlinear and quadratic convergence of affine-scaling interior-point Newton methods for problems with simple bounds without strict complementarity assumption. Math Program, 1999, 86: 615–635
31 Hoffman A J. On approximate solutions of systems of linear inequalities. J Res National Bureau Standards, 1952, 49: 263–265
32 Izmailov A, Solodov M V. Newton-Type Methods for Optimization and Variational Problems. New York: Springer, 2014
33 Kanzow C, Klug A. On affine-scaling interior-point Newton methods for nonlinear minimization with bound constraints. Comput Optim Appl, 2006, 35: 177–197
34 Lescrenier M. Convergence of trust region algorithms for optimization with bounds when strict complementarity does not hold. SIAM J Numer Anal, 1991, 28: 476–495
35 Levitin E S, Polyak B T. Constrained minimization problems. USSR Comput Math Math Phys, 1966, 6: 1–50

36 Lin C-J, Moré J J. Newton's method for large bound-constrained optimization problems. SIAM J Optim, 1999, 9: 1100–1127
37 McCormick G P, Tapia R A. The gradient projection method under mild differentiability conditions. SIAM J Control, 1972, 10: 93–98
38 Moré J J, Toraldo G. On the solution of large quadratic programming problems with bound constraints. SIAM J Optim, 1991, 1: 93–113
39 Polyak B T. The conjugate gradient method in extremal problems. USSR Comp Math Math Phys, 1969, 9: 94–112
40 Robinson S M. Some continuity properties of polyhedral multifunctions. Math Prog Study, 1981, 14: 206–214
41 Robinson S M. Generalized equations and their solutions, part II: applications to nonlinear programming. Math Prog Study, 1982, 19: 200–221
42 Schwartz A, Polak E. Family of projected descent methods for optimization problems with simple bounds. J Optim Theory Appl, 1997, 92: 1–31
43 Ulbrich M, Ulbrich S, Heinkenschloss M. Global convergence of trust-region interior-point algorithms for infinite-dimensional nonconvex minimization subject to pointwise bounds. SIAM J Control Optim, 1999, 37: 731–764
44 Wen Z, Yin W, Goldfarb D, Zhang Y. A fast algorithm for sparse reconstruction based on shrinkage subspace optimization and continuation. SIAM J Sci Comp, 2010, 32: 1832–1857
45 Wen Z, Yin W, Zhang H, Goldfarb D. On the convergence of an active set method for L_1 minimization. Optim Methods Softw, 2012, 27: 1127–1146
46 Wright S J, Nowak R D, Figueiredo M A T. Sparse reconstruction by separable approximation. IEEE Trans Signal Process, 2009, 57: 2479–2493
47 Yang E K, Tolle J W. A class of methods for solving large convex quadratic programs subject to box constraints. Math Program, 1991, 51: 223–228
48 Zhang H, Hager W W. A nonmonotone line search technique and its application to unconstrained optimization. SIAM J Optim, 2004, 14: 1043–1056
49 Zhang Y. Interior-point gradient methods with diagonal-scalings for simple-bound constrained optimization. Tech. Rep. TR04-06. Houston: Rice University, 2004