

1        **A Proximal-Gradient Method for Solving Regularized Optimization**  
 2        **Problems with General Constraints** \*

3        FRANK E. CURTIS<sup>†</sup>, XIAOYI QU<sup>†</sup>, AND DANIEL P. ROBINSON<sup>†</sup>

4        **Abstract.** We propose, analyze, and test a proximal-gradient method for solving regularized  
 5        optimization problems with general constraints. The method employs a decomposition strategy to  
 6        compute trial steps and uses a merit function to determine step acceptance or rejection. Under various  
 7        assumptions, we establish a worst-case iteration complexity result, prove that limit points are first-  
 8        order KKT points, and show that manifold identification and active-set identification properties hold.  
 9        Preliminary numerical experiments on a subset of the CUTEst test problems and sparse canonical  
 10      correlation analysis problems demonstrate the promising performance of our approach.

11        **Key words.** proximal-gradient method, nonlinear optimization, nonconvex optimization, worst-  
 12      case iteration complexity, regularization, composite optimization, constrained optimization

13        **AMS subject classifications.** 49M37, 65K05, 65K10, 65Y20, 68Q25, 90C30, 90C60

14        **1. Introduction.** We consider the constrained optimization problem

15        (1.1)        
$$\min_{x \in \mathbb{R}^n} f(x) + r(x) \text{ subject to (s.t.) } c(x) = 0, x \in \Omega,$$

16        where  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuously differentiable,  $r : \mathbb{R}^n \rightarrow [0, \infty)$  is a nonnegative-  
 17      valued convex function (possibly nonsmooth),  $c : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is continuously differentiable with  $m \leq n$ , and  $\Omega$  is the nonnegative orthant in  $\mathbb{R}^n$  (i.e., the vectors in  $\mathbb{R}^n$  with all nonnegative components). We note that general inequality constraints can  
 18      be converted to the form (1.1) by using slack variables. Thus, problem (1.1) is impor-  
 19      tant to a range of application areas such as data science (e.g., principal component  
 20      analysis [55] and canonical correlation analysis [52, 53]), finance (e.g., portfolio selec-  
 21      tion [1, 14]), signal processing (e.g., sparse blind deconvolution [54] and array beam-  
 22      former design [27, 30]), and image processing (e.g., hyperspectral unmixing [12]).

23        When the constraints in (1.1) are not present, the problem reduces to a nonsmooth  
 24      unconstrained regularized optimization problem, for which proximal-gradient (PG)  
 25      methods and their variants are among the most widely used algorithms [3, 4, 11, 10, 32,  
 26      36]. The basic PG method proceeds by solving a sequence of proximal subproblems.  
 27      Given the  $k$ th iterate  $x_k \in \mathbb{R}^n$  and proximal parameter  $\alpha_k > 0$ , the next iterate  $x_{k+1}$   
 28      is computed as the unique solution to the optimization problem  
 29      Given the  $k$ th iterate  $x_k \in \mathbb{R}^n$  and proximal parameter  $\alpha_k > 0$ , the next iterate  $x_{k+1}$   
 30      is computed as the unique solution to the optimization problem

31        (1.2)        
$$\min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2\alpha_k} \|x - (x_k - \nabla f(x_k))\|_2^2 + r(x) \right\}.$$

32        A notable property of PG methods is that as  $\alpha_k \rightarrow 0$ , the vector  $x_{k+1} - x_k$  con-  
 33      verges to zero. PG methods are also well-known for their *structure identification*  
 34      property [35, 42, 47], whereby the sequence of iterates eventually identifies the mani-  
 35      fold associated with a solution (e.g., the zero-nonzero structure of an optimal solution  
 36      when  $r(x) = \|x\|_1$ ). This property is particularly advantageous in structured opti-  
 37      mization problems for at least three reasons. First, identifying the correct solution  
 38      structure can have significant computational savings. For example, when  $r(x) = \|x\|_1$ ,

\*This material is based upon work supported by the U.S. National Science Foundation, Division of Mathematical Sciences, Computational Mathematics Program under Award Number DMS-1016291.

<sup>†</sup>Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA;  
 E-mail: {frank.e.curtis, xiq322, daniel.p.robinson}@lehigh.edu

39 it is well known that optimal solutions tend to be sparser, and in the context of statistical  
40 modeling sparser solutions offer simpler models that can be employed more efficiently [28, 29]. Second, in certain other applications, the zero-nonzero values of  
41 the variables can have a physical meaning that is lost if the solutions do not have the  
42 true zero-nonzero structure [20, 22, 49]. Third, if the manifold of the solution can  
43 be identified, then one can consider hybrid methods that combine PG calculations  
44 with those of more advanced (usually higher-order) optimization algorithms designed  
45 for *smooth* optimization problems (here restricted to the smooth manifold identified  
46 by the PG iterates). Such an approach aims to exploit local smoothness to achieve  
47 accelerated convergence rates, and has great success in many settings [2, 35, 39].

48 When the regularization function  $r$  is not present in problem (1.1), it reduces to a  
49 traditional nonlinear program. An important concept in the nonlinear programming  
50 literature is *active-set identification*. An algorithm has the active-set identification  
51 property if, under certain reasonable assumptions, it can identify from an iterate near  
52 an optimal solution which inequality constraints are active (i.e., hold at equality)  
53 at that optimal solution. For a comprehensive overview of active-set identification  
54 strategies in nonlinear programming, see [21, 43] and the references therein.

55 Little research has considered the case when the regularization function  $r$  and  
56 nonlinear constraints are present. Two primary challenges arise in this setting. First,  
57 the computation of projections onto the feasible points satisfying  $c(x) = 0$  (or perhaps  
58 the intersection of this region with  $\Omega$ ) is typically computationally intractable. Second,  
59 conventional techniques such as penalty-based methods [17] may fail to preserve the  
60 structure of the solution (see [16, Section 5]), therefore limiting their effectiveness in  
61 this setting. Our work is motivated by the need to address these challenges.

62 **1.1. Related work.** We restrict our attention to work that considers regularized  
63 optimization problems with smooth nonlinear constraints, where both the smooth part  
64 of the objective and the constraints may be nonconvex. Most approaches are penalty-  
65 function-based, where constrained problems are transformed into unconstrained ones  
66 (or ones with simple constraints) by combining the objective function with a penalty  
67 function that measures constraint violation. The resulting subproblems are then typ-  
68 ically solved using the PG method or its variants. Penalty-based methods generally  
69 fall into two main categories: augmented Lagrangian methods and penalty-barrier  
70 methods. Among these, [8, 38, 46] propose inexact augmented Lagrangian meth-  
71 ods and show that an  $\epsilon$ -KKT point can be found within  $\mathcal{O}(\epsilon^{-3})$  iterations under  
72 suitable constraint qualifications. The constraint qualifications in [38, 46] are identi-  
73 cal, whereas [8] uses a slightly different condition, replacing the subdifferential with  
74 the horizon subdifferential. In contrast, the augmented Lagrangian method in [26]  
75 adopts a transversality condition and establishes a better complexity bound of  $\mathcal{O}(\epsilon^{-2})$ .  
76 In [18], an augmented Lagrangian method is proposed for solving regularized problems  
77 with general constraints. The authors use an AM-regularity condition to establish con-  
78 vergence, but no complexity result is provided. To the best of our knowledge, [17] is  
79 the only penalty-barrier approach designed for our problem setting. Instead of assum-  
80 ing any constraint qualification, they directly assume the existence and boundedness  
81 of Lagrange multipliers, which is typically implied by a constraint qualification.

82 Three non-penalty approaches for solving regularized problems with constraints  
83 include [7, 16, 51]. In [51], the authors combine ideas from PG methods and se-  
84 quential quadratic programming methods. In particular, their method formulates a  
85 quadratic approximation to  $f$ , linearizes the constraint function, and keeps the regu-  
86 larizer explicitly in each subproblem. This nonsmooth subproblem is solved using a

88 semi-smooth Newton method. The weakness of this approach is that each subproblem  
 89 is assumed to be feasible and no structure identification result is provided. In [7], a  
 90 feasible proximal-gradient method is proposed that reformulates a nonconvex problem  
 91 into convex surrogate subproblems with quadratic regularization, but it cannot handle  
 92 problems that involve equality constraints due to the infeasibility of each subproblem.  
 93 Our work builds upon on [16], which only considers the equality-constrained case. Al-  
 94 though limited in relevance here, we mention that some work has considered problems  
 95 with only simple bound constraints [5, 34] or only linear constraints [25, 31, 33].

96 **1.2. Contributions.** Our contributions relate to the proposal, analysis, and  
 97 testing of a new PG algorithm for solving problem (1.1), as we now discuss.

- 98 • We propose a new PG method (Algorithm 3.1) for solving problem (1.1). Un-  
 99 like most work in the literature, our method has the following characteristics:  
 100 (i) it uses the regularization function explicitly (as opposed to approximating  
 101 it) when computing the trial step, (ii) it avoids using a penalty function to  
 102 handle the constraints, and (iii) every subproblem is feasible.
- 103 • We establish various convergence results. (i) Without assuming any con-  
 104 straint qualification, we prove that the number of iterations required to re-  
 105 duce a stationarity measure related to minimizing the constraint violation  
 106 below  $\epsilon > 0$  is  $O(\epsilon^{-2})$  (see Theorem 5.8). (ii) Under the linear independence  
 107 constraint qualification (LICQ), we show that all limit points of the iterate  
 108 sequence are first-order KKT points (see Theorem 5.25). (iii) Under a se-  
 109 quential constraint qualification that is stronger than the LICQ, we prove  
 110 that the worst-case iteration complexity needed to reduce a KKT measure  
 111 below  $\epsilon > 0$  is  $O(\epsilon^{-2})$  (see Theorem 5.12). (iv) When strict complementarity  
 112 holds in addition, we prove that our method possesses an optimal active-set  
 113 identification property (see Theorem 5.26). (v) Under partial smoothness of  
 114 the regularization function  $r$  and a certain non-degeneracy assumption, we es-  
 115 tablish a manifold identification property for our method (see Theorem 5.27).
- 116 • We numerically test the performance of our method on CUTEst test prob-  
 117 lems and a sparse canonical correlation analysis problem. In addition, we  
 118 demonstrate the competitive performance of our algorithm by comparing it  
 119 to an augmented Lagrangian approach named Bazinga [18].

120 **1.3. Organization.** In Section 2, we introduce notations and definitions. In  
 121 Section 3, we propose our method as Algorithm 3.1. In Section 4, we derive pre-  
 122 liminary results for the subproblems used in our method, which are critical for the  
 123 theoretical analysis we provide in Section 5. In Section 6, we illustrate our algorithm’s  
 124 performance through numerical tests, and final comments are provided in Section 7.

125 **2. Preliminaries.** Let  $\mathbb{R}$  denote the set of real numbers,  $\mathbb{R}_{\geq 0}$  (resp.,  $\mathbb{R}_{>0}$ ) de-  
 126 note the set of nonnegative (resp., positive) real numbers,  $\mathbb{R}^n$  denote the set of  $n$ -  
 127 dimensional real vectors, and  $\mathbb{R}^{m \times n}$  denote the set of  $m$ -by- $n$ -dimensional real ma-  
 128 trices. The set of natural numbers is  $\mathbb{N} := \{0, 1, 2, \dots\}$ . For a given natural number  
 129  $n \in \mathbb{N}$ , let  $[n] := \{1, \dots, n\}$ . The index sets of active and inactive variables at  $x \in \mathbb{R}^n$   
 130 is  $\mathcal{A}(x) := \{i \in [n] : x_i = 0\}$  and  $\mathcal{I}(x) := \{i \in [n] : x_i \neq 0\}$ , respectively. The  
 131  $\epsilon$ -neighborhood ball of a point  $x \in \mathbb{R}^n$  is  $\mathcal{B}(x, \epsilon) := \{z \in \mathbb{R}^n : \|x - z\|_2 < \epsilon\}$ . Given  
 132 a nonempty set  $\mathcal{C}$  that is either compact, or closed and convex, and a point  $\bar{x} \in \mathbb{R}^n$ ,  
 133 the distance from  $\bar{x}$  to  $\mathcal{C}$  is  $\text{dist}(\bar{x}, \mathcal{C}) := \min_{x \in \mathcal{C}} \|x - \bar{x}\|_2$ .

134 For convenience, we define  $g(x) := \nabla f(x)$  and  $J(x) := \nabla c(x)^T$ . We append a  
 135 natural number as a subscript for a quantity to denote its value during an iteration

136 of an algorithm; i.e., we let  $f_k := f(x_k)$ ,  $g_k := g(x_k)$ ,  $c_k := c(x_k)$ , and  $J_k := J(x_k)$ .

137 We now introduce several key concepts from convex analysis that will be used  
138 throughout the paper. We start with the normal cone [45, Theorem 6.9].

139 **DEFINITION 2.1** (normal cone). *The normal cone of a convex set  $\mathcal{C}$  at  $x \in \mathcal{C}$  is*

140 
$$N_{\mathcal{C}}(x) = \{v \in \mathbb{R}^n : v^T(y - x) \leq 0 \text{ for all } y \in \mathcal{C}\}.$$

141 We define the tangent cone using its polarity with the normal cone [45, Theorem 6.28].

142 **DEFINITION 2.2** (tangent cone). *The tangent cone of a convex set  $\mathcal{C}$  at  $x \in \mathcal{C}$  is*

143 
$$T_{\mathcal{C}}(x) = \{d \in \mathbb{R}^n : v^T d \leq 0 \text{ for all } v \in N_{\mathcal{C}}(x)\}.$$

144 Next, we define the projection onto a closed convex set [6, Proposition 1.1.9].

145 **DEFINITION 2.3** (Projection). *Let  $\mathcal{C} \subseteq \mathbb{R}^n$  be a nonempty closed convex set. The  
146 projection of  $x \in \mathbb{R}^n$  onto  $\mathcal{C}$  is  $\text{Proj}_{\mathcal{C}}(x) := \arg \min_{y \in \mathcal{C}} \|x - y\|_2$ .*

147 Finally, we define the projection of the steepest descent direction of a function  
148 onto the tangent cone [9, Equation (3.1)] associated with  $\Omega$  at a point  $x$ .

149 **DEFINITION 2.4.** *Given a differentiable function  $h : \mathbb{R}^n \rightarrow \mathbb{R}$ , a convex set  $\mathcal{C}$ , and  
150  $x \in \mathcal{C}$ , the projection of the steepest descent direction of  $h$  at  $x$  onto  $T_{\mathcal{C}}(x)$  is*

151 
$$\nabla_{\mathcal{C}} h(x) = \arg \min_{v \in T_{\mathcal{C}}(x)} \|v + \nabla h(x)\|_2 \equiv \text{Proj}_{T_{\mathcal{C}}(x)}(-\nabla h(x)).$$

152 **3. Algorithm Framework.** The algorithm that we propose for solving prob-  
153 lem (1.1) is stated as Algorithm 3.1. Given the  $k$ th iterate  $x_k \in \Omega$ , the  $k$ th proximal  
154 parameter  $\alpha_k$ , and constant  $\kappa_v \in \mathbb{R}_{>0}$ , we first compute a direction  $v_k$  that reduces  
155 linearized infeasibility within  $\Omega$ . In particular, the vector  $v_k$  is computed as an ap-  
156 proximate solution to the bound-constrained trust-region subproblem

157 (3.1) 
$$\min_{v \in \mathbb{R}^n} m_k(v) \text{ s.t. } \|v\|_2 \leq \kappa_v \alpha_k \delta_k, \quad x_k + v \in \Omega \quad \text{with} \quad m_k(v) := \frac{1}{2} \|c_k + J_k v\|_2^2,$$

158 where

159 (3.2) 
$$\delta_k := \|\nabla_{\Omega} \psi(x_k)\|_2 \equiv \|\text{Proj}_{T_{\Omega}(x_k)}(-J_k^T c_k)\|_2 \quad \text{with} \quad \psi(x) := \frac{1}{2} \|c(x)\|_2^2.$$

160 If  $\delta_k = 0$ , then  $v_k \leftarrow 0$  solves (3.1). In this case, if  $\|c_k\|_2 \neq 0$ , we terminate our  
161 algorithm in Line 7 since  $x_k$  is an infeasible stationary point, i.e.,  $x_k$  is infeasible for  
162  $c(x) = 0$  and is a first-order stationary point for the problem

163 (3.3) 
$$\min_{x \in \Omega} \frac{1}{2} \|c(x)\|_2^2.$$

164 If  $\delta_k \neq 0$ , we compute an approximate solution  $v_k$  to (3.1) satisfying

165 (3.4) 
$$\|v_k\|_2 \leq \kappa_v \alpha_k \delta_k, \quad x_k + v_k \in \Omega, \quad \text{and} \quad m_k(v_k) \leq m_k(v_k^c),$$

166 where  $v_k^c$  is a Cauchy point computed using a projected line search along the steepest  
167 descent direction of  $m_k$  at  $v = 0$ . In particular, by defining

168 (3.5) 
$$v_k(\beta) \leftarrow \text{Proj}_{\Omega}(x_k - \beta \nabla m_k(0)) - x_k \equiv \text{Proj}_{\Omega}(x_k - \beta J_k^T c_k) - x_k,$$

169 we define the Cauchy point as

170 (3.6)  $v_k^c := v_k(\beta_k) \equiv \text{Proj}_\Omega(x_k - \beta_k J_k^T c_k) - x_k$

171 where, for some chosen  $\gamma \in (0, 1)$ ,

172 (3.7)  $\beta_k = \gamma^{i_k}$

173 with  $i_k$  being the smallest nonnegative integer such that  $\beta_k$  in (3.7) satisfies

174 (3.8)  $\|v_k(\beta_k)\|_2 \leq \kappa_v \alpha_k \delta_k$  and  $m_k(v_k(\beta_k)) \leq m_k(0) + \eta_m \nabla m_k(0)^T v_k(\beta_k)$

175 for some constant  $\eta_m \in (0, 1)$ . (It follows from Lemma 4.2 later on that this procedure  
176 is well defined.) Note from the definition of  $v_k^c$  (see (3.6) which ensures  $x_k + v_k^c \in \Omega$ )  
177 and (3.8) that  $v_k^c$  itself satisfies the conditions required of  $v_k$  in (3.4).

---

**Algorithm 3.1** PG method for solving problem (1.1)

---

```

1: Input:  $x_0 \in \Omega$ ,  $\{\alpha_0, \tau_{-1}, \kappa_\tau, \kappa_v\} \subset \mathbb{R}_{>0}$ , and  $\{\xi, \eta_\Phi, \sigma_c, \epsilon_\tau, \gamma, \eta_m\} \subset (0, 1)$ 
2: for  $k = 0, 1, 2, \dots$  do
3:   compute  $\delta_k$  in (3.2)
4:   if  $\delta_k = 0$  then
5:     set  $v_k \leftarrow 0$ 
6:     if  $\|c_k\|_2 \neq 0$  then
7:       return  $x_k$  (infeasible stationary point)
8:     end if
9:   else ( $\delta_k \neq 0$ )
10:    compute  $v_k$  as an approximate solution to (3.1) satisfying (3.4)
11:   end if
12:   compute  $u_k$  as the unique solution to subproblem (3.9)
13:   set  $s_k \leftarrow v_k + u_k$ 
14:   if  $\|s_k\|_2/\alpha_k = 0$  then
15:     return  $x_k$  (first-order KKT point for problem (1.1))
16:   end if
17:   compute  $\tau_k$  using (3.10)
18:   if  $\Phi_{\tau_k}(x_k + s_k) - \Phi_{\tau_k}(x_k) \leq -\eta_\Phi \left( \frac{\tau_k}{4\alpha_k} \|s_k\|_2^2 + \sigma_c (\|c_k\|_2 - \|c_k + J_k s_k\|_2) \right)$  then
19:     set  $x_{k+1} \rightarrow x_k + s_k$  and  $\alpha_{k+1} \rightarrow \alpha_k$ 
20:   else
21:     set  $x_{k+1} \rightarrow x_k$  and  $\alpha_{k+1} \rightarrow \xi \alpha_k$ 
22:   end if
23: end for

```

---

178 Next, we compute a direction  $u_k$  that maintains the level of linearized infeasibility  
179 achieved by  $v_k$  while also reducing a model of the objective function. In particular,  
180 we compute  $u_k$  as the unique solution to the strongly convex subproblem

181 (3.9) 
$$\begin{aligned} & \min_{u \in \mathbb{R}^n} g_k^T u + \frac{1}{2\alpha_k} \|u\|_2^2 + \frac{1}{\alpha_k} v_k^T u + r(x_k + v_k + u) \\ & \text{s.t. } J_k u = 0, \quad x_k + v_k + u \in \Omega. \end{aligned}$$

182 Concerning subproblem (3.9), note that  $u = 0$  is feasible and that its solution is unique  
183 since it is a convex optimization problem with a strongly convex objective function.  
184 The overall trial step  $s_k$  is defined as  $s_k = v_k + u_k$ .

To determine whether the trial step  $s_k$  is accepted, we adopt the  $\ell_2$  merit function, which for merit parameter  $\tau \in \mathbb{R}_{>0}$  is defined as

$$\Phi_\tau(x) := \tau(f(x) + r(x)) + \|c(x)\|_2.$$

185 During each iteration, the merit parameter is updated so that  $s_k$  is a descent direction  
186 for the merit function. To ensure that this holds, note that the directional derivative  
187 of  $\Phi_\tau$  at  $x_k$  along  $s_k$ , denoted as  $D_{\Phi_\tau}(x_k, s_k)$ , satisfies (see [16, Lemma 3.3])

$$\begin{aligned} 188 \quad & D_{\Phi_\tau}(x_k, s_k) \\ 189 \quad & \leq \tau(g_k^T s_k + r(x_k + s_k) - r_k) + \|c_k + J_k s_k\|_2 - \|c_k\|_2 \\ 190 \quad & = -\frac{\tau}{2\alpha_k} \|s_k\|_2^2 + \tau \underbrace{(g_k^T s_k + \frac{1}{2\alpha_k} \|s_k\|_2^2 + r(x_k + s_k) - r_k)}_{A_k} + \|c_k + J_k s_k\|_2 - \|c_k\|_2. \end{aligned}$$

191 Next, for a chosen parameter  $\sigma_c \in (0, 1)$ , we set

$$192 \quad \tau_{k,\text{trial}} \leftarrow \begin{cases} \infty & \text{if } A_k \leq 0, \\ \frac{(1-\sigma_c)(\|c_k\|_2 - \|c_k + J_k s_k\|_2)}{g_k^T s_k + \frac{1}{2\alpha_k} \|s_k\|_2^2 + r(x_k + s_k) - r_k} & \text{otherwise,} \end{cases}$$

193 and then set, for some chosen  $\epsilon_\tau \in (0, 1)$ , the value of the  $k$ th merit parameter as

$$194 \quad (3.10) \quad \tau_k \leftarrow \begin{cases} \tau_{k-1} & \text{if } \tau_{k-1} \leq \tau_{k,\text{trial}}, \\ \min\{(1 - \epsilon_\tau)\tau_{k-1}, \tau_{k,\text{trial}}\} & \text{otherwise.} \end{cases}$$

195 This merit parameter update strategy ensures that

$$196 \quad D_{\Phi_{\tau_k}}(x_k, s_k) \leq -\frac{\tau_k}{2\alpha_k} \|s_k\|_2^2 - \sigma_c (\|c_k\|_2 - \|c_k + J_k s_k\|_2),$$

197 meaning that the negative directional derivative is lower bounded by critical measures  
198 of problem (1.1). The  $k$ th iteration is completed by checking whether the merit  
199 function achieves sufficient decrease (see Line 18), and then defining the next iterate  
200 and proximal parameter accordingly. Specifically, if sufficient decrease in the merit  
201 function is achieved, the trial step is accepted (i.e.,  $x_{k+1} \leftarrow x_k + s_k$ ) and the proximal  
202 parameter value is maintained (i.e.,  $\alpha_{k+1} \leftarrow \alpha_k$ ); otherwise, the trial step is rejected  
203 (i.e.,  $x_{k+1} \leftarrow x_k$ ) and the proximal parameter value is decreased (i.e.,  $\alpha_{k+1} \leftarrow \xi \alpha_k$   
204 for some  $\xi \in (0, 1)$ ). This update strategy motivates the definition of the index set

$$205 \quad (3.11) \quad \mathcal{S} := \{k \in \mathbb{N} : x_{k+1} = x_k + s_k\},$$

206 which contains the indices of the successful iterations associated with Algorithm 3.1.  
207 The following assumption is assumed to hold throughout the paper.

208 ASSUMPTION 3.1. *Let  $\mathcal{X} \subseteq \mathbb{R}^n$  be an open convex set containing the iterate se-  
209 quences  $\{x_k\}$  and  $\{x_k + v_k\}$  generated by Algorithm 3.1. The function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$   
210 is bounded over  $\mathcal{X}$ , and its gradient function  $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}$  is Lipschitz continuous  
211 and bounded in norm over  $\mathcal{X}$ . Similarly, for all  $i \in [m]$ , the constraint function  
212  $c_i : \mathbb{R}^n \rightarrow \mathbb{R}$  is bounded over  $\mathcal{X}$ , and its gradient function  $\nabla c_i : \mathbb{R}^n \rightarrow \mathbb{R}$  is Lipschitz  
213 continuous and bounded in norm over  $\mathcal{X}$ . Finally, the function  $r : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$  is  
214 convex, and has bounded subdifferential  $\partial r : \mathbb{R}^n \rightarrow \mathbb{R}^n$  over  $\mathcal{X}$ .*

215 Under Assumption 3.1, there exist constants  $(f_{\inf}, f_{\sup}, \kappa_{\nabla f}, \kappa_{\partial r}, \kappa_c, \kappa_J, L_g, L_J) \in$   
 216  $\mathbb{R} \times \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$  such that for all  $x \in \mathcal{X}$  one has

217 (3.12) 
$$\begin{aligned} f_{\inf} \leq f(x) \leq f_{\sup}, \quad \|\nabla f(x)\|_2 \leq \kappa_{\nabla f}, \quad \|\partial r(x)\|_2 \leq \kappa_{\partial r}, \\ \|c(x)\|_2 \leq \kappa_c, \quad \|\nabla c(x)^T\|_2 \leq \kappa_J, \end{aligned}$$

218 and for all  $(x, \bar{x}) \in \mathcal{X} \times \mathcal{X}$  one has

219 (3.13) 
$$\|\nabla f(x) - \nabla f(\bar{x})\|_2 \leq L_g \|x - \bar{x}\|_2 \text{ and } \|\nabla c(x)^T - \nabla c(\bar{x})^T\|_2 \leq L_J \|x - \bar{x}\|_2.$$

220 **4. Preliminary Properties Related to the Subproblems.** In this section,  
 221 we discuss properties related to the subproblems used in Algorithm 3.1.

222 **4.1. Subproblem (3.1).** In this section, we present properties related to the  
 223 computation of the Cauchy point of subproblem (3.1), following by a final result  
 224 related to the computed feasibility steps. Recall that the Cauchy point is defined  
 225 in (3.6). Our first lemma summarizes properties of  $v_k(\cdot)$  (recall (3.5)).

226 LEMMA 4.1. *Consider  $v_k(\cdot)$  defined in (3.6). For all  $0 < \beta_2 \leq \beta_1$ , it holds that*

227 (4.1a) 
$$\|v_k(\beta_2)\|_2 \leq \|v_k(\beta_1)\|_2 \text{ and}$$
  
 228 (4.1b) 
$$\|v_k(\beta_1)/\beta_1\|_2 \leq \|v_k(\beta_2)/\beta_2\|_2.$$

229 For all  $\beta \in \mathbb{R}_{>0}$  it holds that

230 (4.2a) 
$$-\nabla m_k(0)^T v_k(\beta) \geq \|v_k(\beta)\|_2^2 / \beta \text{ and}$$
  
 231 (4.2b) 
$$\delta_k \equiv \|\nabla_{\Omega} \psi(x_k)\|_2 \geq \|v_k(\beta)/\beta\|_2.$$

232 Finally, the following limit holds:

233 (4.3) 
$$\lim_{\beta \rightarrow 0^+} v_k(\beta)/\beta = \nabla_{\Omega} \psi(x_k).$$

234 *Proof.* Parts (4.1a)–(4.2a) follow from [48, Lemma 2], part (4.3) follows from [40,  
 235 Proposition 2], and part (4.2b) follows by combining (4.3), (4.1b), and (3.2).  $\square$

236 The next result is a special case of [41, Lemma 4.3].

237 LEMMA 4.2. *Suppose that  $\delta_k \neq 0$ . If  $\beta \in \mathbb{R}_{>0}$  satisfies  $m_k(v_k(\beta)) > m_k(0) +$   
 238  $\eta_m \nabla m(0)^T v_k(\beta)$ , then  $\beta \geq (1 - \eta_m) / \|J_k^T J_k\|_2$ .*

239 We now bound the decrease in  $m_k$  by using the argument in [41, Theorem 4.4].

240 LEMMA 4.3. *Suppose that  $\delta_k \neq 0$ . Then, with respect to the constant  $\bar{\kappa}_1 :=$   
 241  $\min\{1, \gamma(1 - \eta_m), \gamma\} \equiv \gamma(1 - \eta_m) \in (0, 1)$ , the Cauchy point  $v_k^c \equiv v_k(\beta_k)$  satisfies*

242 
$$-\nabla m_k(0)^T v_k(\beta_k) \geq \bar{\kappa}_1 \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right] \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2} \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right], \kappa_v \alpha_k \delta_k \right\}.$$

243 Moreover, with respect to the constant  $\kappa_1 := \bar{\kappa}_1 \eta_m \equiv \gamma \eta_m (1 - \eta_m) \in (0, 1)$ , it satisfies

244 
$$m_k(0) - m_k(v_k(\beta_k)) \geq \kappa_1 \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right] \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2} \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right], \kappa_v \alpha_k \delta_k \right\}.$$

245 *Proof.* We begin by proving the first inequality by considering three cases.  
246 **Case 1:**  $\beta_k = 1$ . It follows from (4.2a) and  $\beta_k = 1$  that

$$247 \quad -\nabla m_k(0)^T v_k(\beta_k) \geq \beta_k \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right]^2 = \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right]^2$$

$$248 \quad \geq \frac{\|v_k(\beta_k)\|_2}{\beta_k} \min \left\{ \frac{\|v_k(\beta_k)\|_2}{\beta_k}, \kappa_v \alpha_k \delta_k \right\}.$$

249 Combining this result with  $1/(1 + \|J_k^T J_k\|_2) \leq 1$  shows that the first inequality holds.  
250 **Case 2:**  $\beta_k < 1$  and  $\|v_k(\gamma^{-1} \beta_k)\|_2 \leq \kappa_v \alpha_k \delta_k$ . Since  $\gamma \in (0, 1)$ ,  $\|v_k(\gamma^{-1} \beta_k)\|_2 \leq$   
251  $\kappa_v \alpha_k \delta_k$ , and the step size  $\gamma^{-1} \beta_k$  was not accepted by the search procedure, the suffi-  
252 cient decrease condition must not have held, i.e., it must hold that  $m_k(v_k(\gamma^{-1} \beta_k)) >$   
253  $m_k(0) + \eta_m \nabla m_k(0)^T v_k(\gamma^{-1} \beta_k)$ . Combining this inequality with Lemma 4.2 gives  
254  $\gamma^{-1} \beta_k \geq (1 - \eta_m)/\|J_k^T J_k\|_2$ . Combining this with (4.2a) gives

$$255 \quad -\nabla m_k(0)^T v_k(\beta_k) \geq \beta_k \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right]^2 \geq \gamma \frac{(1 - \eta_m)}{1 + \|J_k^T J_k\|_2} \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right]^2$$

$$256 \quad \geq \gamma (1 - \eta_m) \frac{\|v_k(\beta_k)\|_2}{\beta_k} \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2} \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right], \kappa_v \alpha_k \delta_k \right\}$$

257 so that the first inequality again holds, and completes the proof for this case.

258 **Case 3:**  $\beta_k < 1$  and  $\|v_k(\gamma^{-1} \beta_k)\|_2 > \kappa_v \alpha_k \delta_k$ . It follows from (4.1b) and the fact  
259 that  $\gamma \in (0, 1)$  that  $\frac{\|v_k(\beta_k)\|_2}{\beta_k} \geq \frac{\|v_k(\gamma^{-1} \beta_k)\|_2}{\gamma^{-1} \beta_k}$ . After rearrangement and using the fact  
260 that  $\|v_k(\gamma^{-1} \beta_k)\|_2 > \kappa_v \alpha_k \delta_k$  in this case, we obtain  $\gamma^{-1} \|v_k(\beta_k)\|_2 \geq \|v_k(\gamma^{-1} \beta_k)\|_2 >$   
261  $\kappa_v \alpha_k \delta_k$ , which combined with (4.2a) yields

$$262 \quad -\nabla m_k(0)^T v_k(\beta_k) \geq \|v_k(\beta_k)\|_2 \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right] > \gamma \kappa_v \alpha_k \delta_k \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right]$$

$$263 \quad \geq \gamma \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right] \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2} \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right], \kappa_v \alpha_k \delta_k \right\},$$

264 so that the first inequality again holds, and completes the proof for this case.

265 The second inequality follows from the first inequality and (3.8).  $\square$

266 Combining the previous result with Lemma 4.1 gives new lower bounds.

267 **LEMMA 4.4.** For  $\kappa_1 \in (0, 1]$  in Lemma 4.3, the Cauchy point  $v_k^c \equiv v_k(\beta_k)$  yields

$$268 \quad (4.4a) \quad m_k(0) - m_k(v_k^c) \geq \kappa_1 \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right]^2 \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\}$$

$$269 \quad (4.4b) \quad \geq \kappa_1 \|v_k(1)\|_2^2 \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\}$$

270 and

$$271 \quad (4.5) \quad \|c_k\|_2 - \|c_k + J_k v_k^c\|_2 \geq \frac{\kappa_1}{\kappa_c} \|v_k(1)\|_2^2 \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\}.$$

272 *Proof.* Inequality (4.4a) follows from Lemma 4.3,  $v_k^c = v_k(\beta_k)$ , and (4.2b) with  
273  $\beta = \beta_k$ . Inequality (4.4b) follows from (4.1b) since  $\beta_k \leq 1$ .

274 It follows from (4.4a) that  $\|c_k + J_k v_k^c\|_2 \leq \|c_k\|_2$ . If  $\|c_k\|_2 = 0$ , then (4.5) follows  
275 trivially. Otherwise, it follows from  $\|c_k + J_k v_k^c\|_2 \leq \|c_k\|_2$  that

$$276 \quad (4.6) \quad \begin{aligned} \|c_k\|_2^2 - \|c_k + J_k v_k^c\|_2^2 &= (\|c_k\|_2 + \|c_k + J_k v_k^c\|_2)(\|c_k\|_2 - \|c_k + J_k v_k^c\|_2) \\ &\leq 2\|c_k\|_2(\|c_k\|_2 - \|c_k + J_k v_k^c\|_2). \end{aligned}$$

277 Combining (4.6) and (4.4) we have

$$278 \quad 2\|c_k\|_2(\|c_k\|_2 - \|c_k + J_k v_k^c\|_2) \geq \|c_k\|_2^2 - \|c_k + J_k v_k^c\|_2^2 = 2(m_k(0) - m_k(v_k^c)) \\ 279 \quad \geq 2\kappa_1\|v_k(1)\|_2^2 \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\}.$$

280 Diving both sides by  $2\|c_k\|_2$  and using (3.12) gives (4.5).  $\square$

281 Our next lemma relates the computation of  $v_k$  to the measure  $\delta_k$ . We suspect the  
282 first result is well-known in the literature but we could not find a suitable reference.

283 **LEMMA 4.5.** *The following results hold.*

- 284 (i) *If  $\|v_k(1)\|_2 = 0$ , then  $\delta_k = 0$ .*
- 285 (ii)  *$\|v_k\|_2 = 0$  if and only if  $\delta_k = 0$ .*
- 286 (iii) *If  $\delta_k = 0$ , then  $x_k$  is a first-order KKT point for problem (3.3).*

287 *Proof.* To prove part (i), we suppose that  $\|v_k(1)\|_2 = 0$ . Note that  $0 = \|v_k(1)\|_2 =$   
288  $\|\text{Proj}_\Omega(x_k - J_k^T c_k) - x_k\|_2$  implies that  $\text{Proj}_\Omega(x_k - J_k^T c_k) = x_k$ . Using this fact, we  
289 can apply the projection theorem [6, Proposition 1.1.9] to obtain

$$290 \quad (-J_k^T c_k)^T (z - x_k) = (x_k - J_k^T c_k - x_k)^T (z - x_k) \leq 0 \text{ for all } z \in \Omega,$$

291 which is equivalent to  $-J_k^T c_k \in N_\Omega(x_k)$ . It now follows from Definition 2.2 that

$$292 \quad (4.7) \quad (-J_k^T c_k)^T v \leq 0 \text{ for all } v \in T_\Omega(x_k).$$

293 Using (4.7) and nonnegativity of norms, we find that

$$294 \quad \frac{1}{2}\|v + J_k^T c_k\|_2^2 = \frac{1}{2}(\|v\|_2^2 + 2v^T J_k^T c_k + \|J_k^T c_k\|_2^2) \geq \frac{1}{2}\|J_k^T c_k\|_2^2 \text{ for all } v \in T_\Omega(x_k).$$

It follows from this inequality and  $\frac{1}{2}\|v + J_k^T c_k\|_2^2$  being strongly convex in  $v$  that

$$0 = \arg \min_{v \in T_\Omega(x_k)} \frac{1}{2}\|v + J_k^T c_k\|_2^2 = \arg \min_{v \in T_\Omega(x_k)} \|v + J_k^T c_k\|_2 = \text{Proj}_{T_\Omega(x_k)}(-J_k^T c_k) = \nabla_\Omega(\psi(x_k)).$$

295 It now follows from (3.2) that  $\delta_k = 0$ , which completes the proof of part (i).

296 To prove part (ii), we first observe from Algorithm 3.1 that if  $\delta_k = 0$  then  $v_k = 0$ .  
297 Thus, it remains to prove that if  $v_k = 0$ , then  $\delta_k = 0$ . To do this, let us assume that  
298  $v_k = 0$ . It follows from the third condition in (3.4) and Lemma 4.4 that

$$299 \quad 0 = m_k(0) - m_k(v_k) \geq m_k(0) - m_k(v_k^c) \geq \kappa_1\|v_k(1)\|_2^2 \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\}.$$

300 Since  $\kappa_1$ ,  $\kappa_v$ , and  $\alpha_k$  are strictly positive, it follows that  $\|v_k(1)\|_2 = 0$ . We can combine  
301 this result with part (i) to conclude that  $\delta_k = 0$ , which completes the proof.

302 The proof of part (iii) is provided in [9, Lemma 3.1(c)].  $\square$

303     **4.2. Subproblem (3.9).** With respect to subproblem (3.9), we recall that  $u = 0$   
 304 is feasible, the constraints are linear (meaning that the feasible region is convex and  
 305 that a constraint qualification holds), and the objective function is strongly convex.  
 306 Therefore, the unique solution  $u_k$  to subproblem (3.9) satisfies, for some  $g_{r,k} \in \partial r(x_k +$   
 307  $v_k + u_k)$ ,  $y_k \in \mathbb{R}^m$ , and  $z_k \in \mathbb{R}^n$ , the following conditions:

308 (4.8a)      $g_k + \frac{1}{\alpha_k} u_k + \frac{1}{\alpha_k} v_k + g_{r,k} + J_k^T y_k + z_k = 0,$

309 (4.8b)      $J_k u_k = 0, \text{ and}$

310 (4.8c)      $\| \min\{x_k + v_k + u_k, -z_k\} \|_2 = 0,$

311 where the minimum of two vectors is taken componentwise. These conditions charac-  
 312 terize  $u_k$  and will play a critical role in the analysis of Section 5. In particular, they  
 313 allow us to establish the following bound on the size of the trial step.

314     LEMMA 4.6. *The trial step  $s_k$  satisfies  $\|s_k\|_2 \geq \|\min\{x_k, -z_k\}\|_2$ .*

315     *Proof.* It follows from  $s_k = v_k + u_k$  and (4.8) that

316 (4.9)      $-\frac{1}{\alpha_k} s_k = g_k + g_{r,k} + J_k^T y_k + z_k \text{ and } \|\min\{x_k + s_k, -z_k\}\|_2 = 0.$

317 The latter equality and min-inequalities give, for each  $i \in \{1, 2, \dots, n\}$ , that

318      $0 = \min\{[x_k + s_k]_i, -[z_k]_i\} \geq \min\{[x_k]_i, -[z_k]_i\} + \min\{[s_k]_i, 0\}.$

319 Combining this inequality with  $\min\{[x_k]_i, -[z_k]_i\} \geq 0$  gives  $0 \leq \min\{[x_k]_i, -[z_k]_i\} \leq$   
 320  $-\min\{[s_k]_i, 0\}$ . It follows from this inequality that

321     
$$\begin{aligned} \|\min\{x_k, -z_k\}\|_2^2 &= \sum_{i=1}^n |\min\{[x_k]_i, -[z_k]_i\}|^2 \\ 322 &\leq \sum_{i=1}^n |\min\{[s_k]_i, 0\}|^2 \leq \sum_{i=1}^n |[s_k]_i|^2 = \|s_k\|_2^2. \end{aligned}$$

323 Taking the square-root of both sides of this inequality completes the proof.  $\square$

324     **5. Analysis.** In this section, we present a complete convergence analysis for  
 325 Algorithm 3.1 in both the finite termination case and infinite iteration case.

326     **5.1. Finite termination.** Our first result shows that the solutions to our sub-  
 327 problems that define the trial step are both zero precisely when the trial step is zero.

328     LEMMA 5.1.  *$s_k = 0$  if and only if  $v_k = u_k = 0$ .*

329     *Proof.* Since  $s_k = v_k + u_k$ , it follows that if  $v_k = u_k = 0$ , then  $s_k = 0$ . Thus,  
 330 it remains to prove that if  $s_k = 0$ , then  $v_k = u_k = 0$ . For a proof by contradiction,  
 331 suppose that  $s_k = 0$  and  $v_k \neq 0$ . It follows from Lemma 4.5(i)(ii) that  $v_k(1) \neq 0$ , so  
 332 that Lemma 4.4 gives  $v_k^c \neq 0$ . We may now combine this result with (4.2a) to obtain

333      $c_k^T J_k v_k^c = (J_k^T c_k)^T v_k^c = \nabla m_k(0)^T v_k^c \leq -\|v_k^c\|_2^2 / \beta_k < 0,$

334 which implies that  $J_k v_k^c \neq 0$ , i.e., that  $v_k^c$  is not in the nullspace of  $J_k$ . At the  
 335 same time, we know from (4.8b) that  $u_k$  is in the nullspace of  $J_k$ . The previous two  
 336 statements cannot both be true since  $s_k = v_k + u_k = 0$  implies that  $v_k = -u_k$ , which  
 337 is a contradiction. Therefore, we must conclude that  $v_k = 0$ . Combining this result  
 338 with  $s_k = v_k + u_k = 0$  shows that  $u_k = 0$ , and completes the proof.  $\square$

339 We can now state our finite termination results for Algorithm 3.1.

340 **THEOREM 5.2.** *The following finite termination results hold for Algorithm 3.1.*

341 (i) *If Algorithm 3.1 terminates at Line 7, then  $x_k$  is an infeasible stationary*  
 342 *point, i.e.,  $x_k$  is a first-order KKT point for problem (3.3) and  $\|c_k\|_2 \neq 0$ .*  
 343 (ii) *If Algorithm 3.1 terminates at Line 15, then  $x_k$  is a first-order KKT point*  
 344 *for problem (1.1).*

345 *Proof.* We first prove part (i). If Algorithm 3.1 terminates at Line 7, then it  
 346 follows from Lines 4 and 6 that  $\delta_k = 0$  and  $\|c_k\|_2 \neq 0$ . It now follows from  $\delta_k = 0$  and  
 347 Lemma 4.5(iii) that  $x_k$  is a first-order KKT point for problem (3.3), as claimed.

348 For part (ii), we know that if Algorithm 3.1 terminates in Line 15 then  $s_k = 0$ ,  
 349 which from Lemma 5.1 implies that  $u_k = v_k = 0$ , and then Lemma 4.5(ii) implies  
 350 that  $\delta_k = 0$ . Since termination did not occur in Line 7 of Algorithm 3.1, we know  
 351 that  $\|c_k\|_2 = 0$ . It follows from  $v_k = u_k = 0$  and (4.8) that there exists  $g_{r,k} \in \partial r(x_k)$ ,  
 352  $y_k \in \mathbb{R}^m$ , and  $z_k \in \mathbb{R}^n$  satisfying  $g_k + g_{r,k} + J_k^T y_k + z_k = 0$  and  $\|\min\{x_k, -z_k\}\|_2 = 0$ .  
 353 These equations and  $\|c_k\|_2 = 0$  show that  $x_k$  is a first-order KKT point for (1.1).  $\square$

354 **5.2. Infinite iterations.** We now consider the scenario where finite termination  
 355 does not occur, meaning that Algorithm 3.1 performs an infinite number of iterations.

356 **5.2.1. Analysis under no constraint qualification.** In this section, we an-  
 357 alyze properties of the iterate sequence  $\{x_k\}$  generated by Algorithm 3.1 when no  
 358 constraint qualification is assumed to hold. The key metric we consider is

359 (5.1) 
$$\bar{\chi}_k := \max \{ \|g_k + g_{r,k} + J_k^T y_k + z_k\|_2, \|v_k(1)\|_2, \|\max\{x_k, -z_k\}\|_2 \},$$

360 where  $g_{r,k} \in \mathbb{R}^n$ ,  $y_k \in \mathbb{R}^m$ , and  $z_k \in \mathbb{R}^n$  are defined as those quantities satisfying (4.8).  
 361 The first quantity in the max is a measure of stationarity for problem (1.1), the second  
 362 quantity is a stationarity measure for problem (3.3), and the third quantity measures  
 363 feasibility with respect to  $x_k \in \Omega$ , the sign of the Lagrange multiplier estimate  $z_k$ , and  
 364 complementarity. In particular, we emphasize that  $\|v_k(1)\|_2$  is used here in place of  
 365  $\|c_k\|_2$  since a constraint qualification is not assumed to hold in this section, meaning  
 366 that it is possible that the iterates do not converge toward feasibility.

367 Our first result gives a uniform upper bound on the sequence  $\{\delta_k\}$  defined in (3.2).

368 **LEMMA 5.3.** *For all iterations  $k \in \mathbb{N}$ , we have that*

369 (5.2) 
$$\delta_k \equiv \|\nabla_{\Omega} \psi(x_k)\|_2 \leq 2\kappa_J \|c_k\|_2 \leq 2\kappa_J \kappa_c.$$

370 *Proof.* Recall that  $\nabla_{\Omega} \psi(x_k) = \arg \min \{ \|v + J_k^T c_k\|_2 : v \in T_{\Omega}(x_k) \}$ . It follows  
 371 from this fact, the triangle inequality, and  $0 \in T_{\Omega}(x_k)$  that

372 
$$\|\nabla_{\Omega} \psi(x_k)\|_2 - \|J_k^T c_k\|_2 \leq \|\nabla_{\Omega} \psi(x_k) + J_k^T c_k\|_2 \leq \|J_k^T c_k\|_2.$$

373 It follows from this inequality, how  $\delta_k$  is defined in (3.2), and Assumption 3.1 that  
 374  $\delta_k \equiv \|\nabla_{\Omega} \psi(x_k)\|_2 \leq 2\|J_k^T c_k\|_2 \leq 2\kappa_J \|c_k\|_2 \leq 2\kappa_J \kappa_c$ , which completes the proof.  $\square$

375 We can now prove an upper bound on  $A_k$  that is defined for  $\tau_{k,\text{trial}}$ .

376 **LEMMA 5.4.** *For all  $k \in \mathbb{N}$ , we have that*

377 
$$g_k^T s_k + \frac{1}{2\alpha_k} \|s_k\|_2^2 + r(x_k + s_k) - r_k \leq 2(\kappa_{\nabla f} + \kappa_{\partial r}) \kappa_v \kappa_J \alpha_k \|c_k\|_2 + 2\kappa_v^2 \kappa_J^2 \kappa_c \alpha_k \|c_k\|_2.$$

378 *Proof.* By convexity of  $r$ , we know that

379 (5.3) 
$$r(x_k + v_k) - r_k \leq (g_{r,k}^v)^T v_k \text{ for all } g_{r,k}^v \in \partial r(x_k + v_k).$$

380 It now follows that

$$\begin{aligned}
381 \quad & g_k^T s_k + \frac{1}{2\alpha_k} \|s_k\|_2^2 + r(x_k + s_k) - r_k \\
382 \quad & \stackrel{(i)}{\leq} g_k^T v_k + \frac{1}{2\alpha_k} \|v_k\|_2^2 + r(x_k + v_k) - r_k \\
383 \quad & \stackrel{(ii)}{\leq} g_k^T v_k + \frac{1}{2\alpha_k} \|v_k\|_2^2 + (g_{r,k}^v)^T v_k \\
384 \quad & \stackrel{(iii)}{\leq} (\|g_k\|_2 + \|g_{r,k}^v\|_2) \|v_k\|_2 + \frac{1}{2\alpha_k} \|v_k\|_2^2 \\
385 \quad & \stackrel{(iv)}{\leq} (\|g_k\|_2 + \|g_{r,k}^v\|_2) \kappa_v \alpha_k \delta_k + \frac{1}{2\alpha_k} \kappa_v^2 \alpha_k^2 \delta_k^2 \\
386 \quad & \stackrel{(v)}{=} (\|g_k\|_2 + \|g_{r,k}^v\|_2) \kappa_v \alpha_k \delta_k + \frac{1}{2} \kappa_v^2 \alpha_k \delta_k^2 \\
387 \quad & \stackrel{(vi)}{\leq} (\|g_k\|_2 + \|g_{r,k}^v\|_2) 2\kappa_v \alpha_k \kappa_J \|c_k\|_2 + 2\kappa_v^2 \alpha_k \kappa_J^2 \kappa_c \|c_k\|_2 \\
388 \quad & \stackrel{(vii)}{\leq} (\kappa_{\nabla f} + \kappa_{\partial r}) 2\kappa_v \kappa_J \alpha_k \|c_k\|_2 + 2\kappa_v^2 \kappa_J^2 \kappa_c \alpha_k \|c_k\|_2,
\end{aligned}$$

389 where (i) follows from substituting  $s_k = v_k + u_k$  and using the fact that  $u_k = 0$  is a  
390 feasible solution to the tangential subproblem (3.9), (ii) follows from (5.3), (iii) follows  
391 from the Cauchy-Schwartz inequality, (iv) follows from  $\|v_k\|_2 \leq \kappa_v \alpha_k \delta_k$  in (3.4), (v)  
392 follows from canceling an  $\alpha_k$  from the second term, (vi) follows from Lemma 5.3  
393 and (3.12), and (vii) follows from (3.12). This completes the proof.  $\square$

394 The first part of the next lemma establishes that the merit parameter never needs  
395 to be decreased for any iteration  $k \in \mathbb{N}$  such that  $v_k(1) = 0$ . On the other hand, for  
396 all  $k \in \mathbb{N}$  satisfying  $v_k(1) \neq 0$ , the second part of the lemma provides a lower bound  
397 on how small the previous merit parameter  $\tau_{k-1}$  could have been when decreased.

398 **LEMMA 5.5.** *The following merit parameter update results hold.*

399 (i) *For each  $k \in \mathbb{N} \setminus \{0\}$ , if  $v_k(1) = 0$ , then  $\tau_{k,\text{trial}} = \infty$  and  $\tau_k \leftarrow \tau_{k-1}$ .*  
400 (ii) *There exists a constant  $\epsilon_\tau > 0$  such that, for all  $k \in \mathbb{N}$  satisfying  $\|v_k(1)\|_2 \neq 0$   
401 and  $\tau_k < \tau_{k-1}$ , it holds that  $\tau_{k-1} \geq \epsilon_\tau \|v_k(1)\|_2^2$ .*

402 *Proof.* We first prove part (i). To this end, first observe that  $v_k(1) = 0$  and  
403 Lemma 4.5(i) imply that  $\delta_k = 0$ , and therefore  $v_k = 0$  holds as a consequence of  
404 Lemma 4.5(ii). Next, since  $u = 0$  is feasible for subproblem (3.9) we know that

$$g_k^T u_k + \frac{1}{2\alpha_k} \|u_k\|_2^2 + \frac{1}{\alpha_k} v_k^T u_k + r(x_k + v_k + u_k) \leq r(x_k + v_k),$$

402 which may be combined with  $v_k = 0$  to obtain  $g_k^T s_k + \frac{1}{2\alpha_k} \|s_k\|_2^2 + r(x_k + s_k) \leq r(x_k)$ .  
403 This inequality and the definition of  $\tau_{k,\text{trial}}$  gives  $\tau_{k,\text{trial}} = \infty$ , so that  $\tau_k \leftarrow \tau_{k-1}$ .

404 Next, we prove part (ii). It follows from the merit parameter update rule (3.10),  
405  $J_k u_k = 0$  (see (4.8b)), the third condition in (3.4), (4.5), (3.12), Lemma 5.4, and  
406 monotonicity of the proximal parameter sequence  $\{\alpha_k\}$  that if  $\tau_k < \tau_{k-1}$ , then

$$\begin{aligned}
407 \quad & \tau_{k-1} > \frac{(1 - \sigma_c)(\|c_k\|_2 - \|c_k + J_k v_k\|_2)}{g_k^T s_k + \frac{1}{2\alpha_k} \|s_k\|_2^2 + r(x_k + s_k) - r_k} \\
408 \quad & \geq \frac{(1 - \sigma_c)(\|c_k\|_2 - \|c_k + J_k v_k^c\|_2)}{g_k^T s_k + \frac{1}{2\alpha_k} \|s_k\|_2^2 + r(x_k + s_k) - r_k} \\
409 \quad & \geq \frac{(1 - \sigma_c) \frac{\kappa_1}{\kappa_c} \|v_k(1)\|_2^2 \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\}}{2(\kappa_{\nabla f} + \kappa_{\partial r}) \kappa_v \kappa_J \alpha_k \|c_k\|_2 + 2\kappa_v^2 \kappa_J^2 \kappa_c \alpha_k \|c_k\|_2}
\end{aligned}$$

410 
$$\geq \frac{(1 - \sigma_c)\kappa_1\|v_k(1)\|_2^2 \min\left\{\frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k\right\}}{2(\kappa_{\nabla f} + \kappa_{\partial r})\kappa_v \kappa_J \kappa_c^2 \alpha_k + 2\kappa_v^2 \kappa_J^2 \kappa_c^2 \alpha_k} \geq \epsilon_\tau \|v_k(1)\|_2^2,$$

411 where  $\epsilon_\tau := \frac{(1 - \sigma_c)\kappa_1 \min\left\{\frac{1}{1 + \kappa_J^2 \alpha_0}, \kappa_v\right\}}{2(\kappa_{\nabla f} + \kappa_{\partial r})\kappa_v \kappa_J \kappa_c^2 + 2\kappa_v^2 \kappa_J^2 \kappa_c^2} > 0$ , thus completing the proof.  $\square$

412 Next, under the assumption that the merit parameter sequence stays bounded  
413 away from zero, we give a positive lower bound on  $\{\alpha_k\}$ .

414 LEMMA 5.6. *Assume that there exists  $\tau_{\min} > 0$  such that  $\tau_k \geq \tau_{\min}$  for all  $k \in \mathbb{N}$ .  
415 If  $\alpha_k \leq \frac{\tau_{\min}}{2(\tau_{\min} L_g + L_J)}$ , then  $k \in \mathcal{S}$ . Thus, for all  $k \in \mathbb{N}$ ,*

416 (5.4) 
$$\alpha_k \geq \alpha_{\min} := \min\{\alpha_0, \frac{\xi \tau_{\min}}{2(\tau_{\min} L_g + L_J)}\} > 0$$

417 and a bound on the number of unsuccessful iterations is given by

418 (5.5) 
$$|\{k \in \mathbb{N} : x_k \notin \mathcal{S}\}| \leq \max\left(0, \left\lceil \frac{\log\left(\frac{\tau_{\min}}{2\alpha_0(\tau_{\min} L_g + L_J)}\right)}{\log(\xi)} \right\rceil\right).$$

419 *Proof.* It follows from (3.13) and the merit parameter update rule (3.10) that

420 (5.6) 
$$\begin{aligned} & \Phi_{\tau_k}(x_k + s_k) - \Phi_{\tau_k}(x_k) \\ &= \tau_k(f(x_k + s_k) + r(x_k + s_k)) + \|c(x_k + s_k)\|_2 - \tau_k(f_k + r_k) - \|c_k\|_2. \\ &\leq \tau_k g_k^T s_k + \tau_k(r(x_k + s_k) - r_k) + \|c_k + J_k s_k\|_2 - \|c_k\|_2 + \frac{1}{2}(\tau_k L_g + L_J) \|s_k\|_2^2 \\ &\leq -\frac{\tau_k}{4\alpha_k} \|s_k\|_2^2 - \sigma_c(\|c_k\|_2 - \|c_k + J_k s_k\|_2) + \frac{1}{2}(-\frac{\tau_k}{2\alpha_k} + \tau_k L_g + L_J) \|s_k\|_2^2. \end{aligned}$$

421 Suppose that  $k \in \mathbb{N}$  satisfies  $\alpha_k \leq \frac{\tau_{\min}}{2(\tau_{\min} L_g + L_J)}$ . It follows from the fact that  $\frac{\tau}{2(\tau L_g + L_J)}$   
422 is a monotonically increasing function on the nonnegative real line as a function  
423 of  $\tau$  that  $\alpha_k \leq \frac{\tau_{\min}}{2(\tau_{\min} L_g + L_J)} \leq \frac{\tau_k}{2(\tau_k L_g + L_J)}$ , which after rearrangement shows that  
424  $-\frac{\tau_k}{2\alpha_k} + \tau_k L_g + L_J \leq 0$ . The previous inequality,  $\|s_k\|_2 \neq 0$  (since finite termination  
425 does not occur), (4.5),  $\|c_k + J_k v_k\|_2 \leq \|c_k + J_k v_k^c\|_2$ ,  $J_k u_k = 0$ , and  $\eta_\Phi \in (0, 1)$  give

426 
$$(1 - \eta_\Phi)(\frac{\tau_k}{4\alpha_k} \|s_k\|_2^2 + \sigma_c(\|c_k\|_2 - \|c_k + J_k s_k\|_2)) > 0 \geq \frac{1}{2}(-\frac{\tau_k}{2\alpha_k} + \tau_k L_g + L_J) \|s_k\|_2^2.$$

427 Combining this inequality with (5.6) shows that  $k \in \mathcal{S}$ , as claimed. This result and  
428 the update strategy for the proximal parameter  $\alpha_k$  ensures that the bound in (5.4)  
429 holds. Finally, the first result we proved in this lemma and the update strategy  
430 for  $\{\alpha_k\}$  shows that the maximum number of unsuccessful iterations is the smallest  
431 nonnegative integer  $n_u$  such that  $\xi^{n_u} \alpha_0 \leq \frac{\tau_{\min}}{2(\tau_{\min} L_g + L_J)}$ , which gives the final result.  $\square$

432 It will be convenient for our analysis to define the shifted merit function

433 (5.7) 
$$\bar{\Phi}_\tau(x) := \tau(f(x) - f_{\inf} + r(x)) + \|c(x)\|_2,$$

434 where  $f_{\inf}$  is defined in (3.12). We stress that the (typically) unknown value  $f_{\inf}$  is  
435 never used in the algorithm statement or its implementation, only in our analysis.

436 LEMMA 5.7. *The following properties hold for the shifted merit function.*

437 (i) *For all  $\{x, y\} \subset \mathbb{R}^n$  and  $\tau \in \mathbb{R}_{>0}$ , it holds that  $\bar{\Phi}_\tau(x) - \bar{\Phi}_\tau(y) = \Phi_\tau(x) - \Phi_\tau(y)$ .*  
438 (ii) *For all  $x \in \mathbb{R}^n$  and  $0 < \tau_2 \leq \tau_1$ , it holds that  $\bar{\Phi}_{\tau_2}(x) \leq \bar{\Phi}_{\tau_1}(x)$ .*

439 (iii) The sequence  $\{\bar{\Phi}_{\tau_k}(x_k)\}$  is monotonically decreasing.

440 *Proof.* See [16, Lemma 3.14] for a proof.  $\square$

441 We can now state our main convergence result for this section.

442 THEOREM 5.8. *Let Assumption 3.1 hold. One of the following two cases occurs.*

443 (i) *There exists  $\tau_{\min} > 0$  such that  $\tau_k \geq \tau_{\min}$  for all  $k \in \mathbb{N}$ . In this case, the*  
 444 *following hold: (a)  $\alpha_k \geq \alpha_{\min} := \min\{\alpha_0, \frac{\xi \tau_{\min}}{2(\tau_{\min} L_g + L_J)}\}$  for all  $k \in \mathbb{N}$ ; (b) If*  
 445  *$\{k_1, k_2\} \subset \mathbb{N}$  are two iterations with  $k_1 < k_2$  such that  $k \in \mathcal{S}$  and  $\bar{\chi}_k > \epsilon$  for*  
 446 *all iterations  $k_1 \leq k < k_2$ , then it follows that*

$$447 (5.8) \quad k_2 - k_1 \leq \left\lceil \frac{\tau_0(f(x_0) + r(x_0) - f_{\inf}) + \|c(x_0)\|_2}{\bar{\kappa}_\Phi \epsilon^2} \right\rceil$$

448 with  $\bar{\kappa}_\Phi = \eta_\Phi \min \left\{ \frac{\tau_{\min} \alpha_{\min}}{8}, \frac{\tau_{\min}}{8\alpha_0}, \frac{\sigma_c \kappa_1}{\kappa_c} \min \left\{ \frac{1}{1+\kappa_J^2}, \kappa_v \alpha_{\min} \right\} \right\}$ ; and (c) for any  
 449 given  $\epsilon > 0$ , the maximum number of iterations before  $\bar{\chi}_k \leq \epsilon$  is

$$450 \quad \left( \max \left\{ 0, \left\lceil \frac{\log \left( \frac{\tau_{\min}}{2\alpha_0(\tau_{\min} L_g + L_J)} \right)}{\log(\xi)} \right\rceil \right\} + 1 \right) \left\lceil \frac{\tau_0(f(x_0) - f_{\inf} + r(x_0)) + \|c(x_0)\|_2}{\bar{\kappa}_\Phi \epsilon^2} \right\rceil.$$

451 (ii) *The merit parameter values converge to zero, i.e.,  $\lim_{k \rightarrow \infty} \tau_k = 0$ . In this*  
 452 *case, there exists a subsequence  $\mathcal{K} \subseteq \mathbb{N}$  such that  $\lim_{k \in \mathcal{K}} \|v_k(1)\|_2 = 0$ .*

453 *Proof.* To prove part (i), let us assume there exists  $\tau_{\min} > 0$  such that  $\tau_k \geq \tau_{\min}$   
 454 for all  $k \in \mathbb{N}$ . Using this fact, Lemma 5.6 ensures that both (5.4) and (5.5) hold.  
 455 Since (5.4) holds, part (i)(a) is proved. To prove part (i)(b), let  $\{k_1, k_2\}$  be as in  
 456 the statement of the theorem. Then, for all  $k \in \mathcal{S}$  and  $k_1 \leq k < k_2$ , it follows from  
 457 Lemma 5.7(i)–(ii),  $k \in \mathcal{S}$ , (3.12),  $J_k u_k = 0$ , (4.5), and Lemma 5.6 that

$$458 (5.9) \quad \begin{aligned} \bar{\Phi}_{\tau_k}(x_k) - \bar{\Phi}_{\tau_{k+1}}(x_{k+1}) &\geq \bar{\Phi}_{\tau_k}(x_k) - \bar{\Phi}_{\tau_k}(x_{k+1}) = \Phi_{\tau_k}(x_k) - \Phi_{\tau_k}(x_{k+1}) \\ &\geq \eta_\Phi \left( \frac{\tau_k}{4\alpha_k} \|s_k\|_2^2 + \sigma_c (\|c_k\|_2 - \|c_k + J_k s_k\|_2) \right) \\ &\geq \eta_\Phi \left[ \frac{\tau_k \alpha_k}{4} \left( \frac{\|s_k\|_2}{\alpha_k} \right)^2 + \frac{\sigma_c \kappa_1}{\kappa_c} \|v_k(1)\|_2^2 \min \left\{ \frac{1}{1+\|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\} \right] \\ &= \eta_\Phi \left[ \frac{\tau_k \alpha_k}{8} \left( \frac{\|s_k\|_2}{\alpha_k} \right)^2 + \frac{\tau_k \|s_k\|_2^2}{8\alpha_k} + \frac{\sigma_c \kappa_1}{\kappa_c} \|v_k(1)\|_2^2 \min \left\{ \frac{1}{1+\|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\} \right]. \end{aligned}$$

459 Lemma 4.6, (5.9), (4.8), (5.4), and  $\tau_k \geq \tau_{\min}$  and  $\alpha_k \leq \alpha_0$  for all  $k \in \mathbb{N}$  give

$$460 \quad \begin{aligned} \bar{\Phi}_{\tau_k}(x_k) - \bar{\Phi}_{\tau_{k+1}}(x_{k+1}) &\geq \eta_\Phi \left[ \frac{\tau_k \alpha_k}{8} \|g_k + g_{r,k} + J_k^T y_k + z_k\|_2^2 + \frac{\tau_k}{8\alpha_k} \|\min\{x_k, -z_k\}\|_2^2 \right. \\ 461 &\quad \left. + \frac{\sigma_c \kappa_1}{\kappa_c} \|v_k(1)\|_2^2 \min \left\{ \frac{1}{1+\|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\} \right] \\ 462 &\geq \eta_\Phi \left[ \frac{\tau_{\min} \alpha_{\min}}{8} \|g_k + g_{r,k} + J_k^T y_k + z_k\|_2^2 + \frac{\tau_{\min}}{8\alpha_0} \|\min\{x_k, -z_k\}\|_2^2 \right. \\ 463 &\quad \left. + \frac{\sigma_c \kappa_1}{\kappa_c} \|v_k(1)\|_2^2 \min \left\{ \frac{1}{1+\kappa_J^2}, \kappa_v \alpha_{\min} \right\} \right] \\ 464 &\geq \bar{\kappa}_\Phi \bar{\chi}_k^2 \end{aligned}$$

465 where  $\bar{\kappa}_\Phi$  is defined in the statement of the current theorem. Using this inequality,  
 466 Lemma 5.7(iii), and nonnegativity of  $\bar{\Phi}_\tau$  for all  $\tau \in \mathbb{R}_{>0}$ , we find that

$$468 \quad \bar{\Phi}_{\tau_0}(x_0) \geq \bar{\Phi}_{\tau_{k_1}}(x_{k_1}) \geq \bar{\Phi}_{\tau_{k_1}}(x_{k_1}) - \bar{\Phi}_{\tau_{k_2}}(x_{k_2})$$

469

$$= \sum_{k=k_1}^{k_2-1} (\bar{\Phi}_{\tau_k}(x_k) - \bar{\Phi}_{\tau_{k+1}}(x_{k+1})) \geq \sum_{k=k_1}^{k_2-1} \bar{\kappa}_\Phi \bar{\chi}_k^2,$$

470 which may be combined with  $\bar{\chi}_k > \epsilon$  for all  $k_1 \leq k \leq k_2$  to conclude that  $\bar{\Phi}_{\tau_0}(x_0) \geq$   
471  $(k_2 - k_1)\bar{\kappa}_\Phi \epsilon^2$ , from which (5.8) follows. The result (i)(c), namely the claimed upper  
472 bound on the maximum iterations before  $\bar{\chi}_k \leq \epsilon$ , follows from what we just proved  
473 and the fact that maximum number of unsuccessful iterations is bounded as in (5.5).

474 We prove part (ii) by contradiction. Thus, suppose that there exists  $\epsilon \in \mathbb{R}_{>0}$  and  
475  $\bar{k}_1 \in \mathbb{N}$  such that  $\|v_k(1)\|_2 \geq \epsilon$  for all  $k \geq \bar{k}_1$ . It then follows from Lemma 5.5 that  
476 there exists  $\tau_{\min} \in \mathbb{R}_{>0}$  such that  $\tau_k \geq \tau_{\min}$  for all  $k \in \mathbb{N}$ , which is a contradiction.  $\square$

**5.2.2. Analysis under a sequential constraint qualification.** In this section, we assume that a sequential constraint qualification holds (all results from Section 5.2.1 still hold). To state this assumption, we define the index set of active variables after taking the Cauchy step  $v_k^c$  as

$$\mathcal{A}_k^v := \mathcal{A}(x_k + v_k^c) \equiv \{i \in [n] : [x_k + v_k^c]_i = 0\}.$$

477 We can now formally state the assumption we make throughout this section.

478 **ASSUMPTION 5.1.** *The matrix  $[J_k^T, I_{\mathcal{A}_k^v}^T]^T$  has full row rank and its smallest singular value is uniformly bounded away from zero for all  $k \in \mathbb{N}$ , where  $I_{\mathcal{A}_k^v}$  denotes the subset of rows of the identity matrix that correspond to the elements in  $\mathcal{A}_k^v$ , i.e., there exists  $\sigma_{\min} \in \mathbb{R}_{>0}$  such that  $\sigma_{\min}([J_k^T, I_{\mathcal{A}_k^v}^T]^T) \geq \sigma_{\min}$  for all  $k \in \mathbb{N}$  with  $\sigma_{\min}(A)$  denoting the smallest singular value of a matrix  $A$ .*

483 Under the above assumption, our aim is to prove a worst-case iteration complexity  
484 result for Algorithm 3.1. Our result uses the KKT-residual measure

485 (5.10)  $\chi_k := \max \{ \|g_k + g_{r,k} + J_k^T y_k + z_k\|_2, \|c_k\|_2, \|\min\{x_k, -z_k\}\|_2 \}.$

486 Note that (5.10) differs from the definition of  $\bar{\chi}_k$  in (5.1) by using the measure  $\|c_k\|_2$   
487 instead of  $\|v_k(1)\|_2$ , which is reasonable because of the constraint qualification.

488 We begin by establishing a key connection between  $\|v_k(\beta_k)\|_2$  and  $\|c_k\|_2$ .

489 **LEMMA 5.9.** *For all  $k \in \mathbb{N}$ , it holds that  $\|v_k(\beta_k)\|_2 / \beta_k \geq \sigma_{\min} \|c_k\|_2$ .*

490 *Proof.* Let us define the vector  $w_k \in \mathbb{R}^n$  componentwise as

491 (5.11)  $[w_k]_i = \begin{cases} 0 & i \in [n] \setminus \mathcal{A}_k^v, \\ -[J_k^T c_k]_i - [v_k(\beta_k)]_i / \beta_k & i \in \mathcal{A}_k^v. \end{cases}$

492 We claim that the following holds:

493 (5.12)  $\text{Proj}_\Omega(x_k - \beta_k J_k^T c_k) - x_k = -\beta_k J_k^T c_k - \beta_k w_k,$

494 which we verify by considering its coordinates. If  $i \in \mathcal{A}_k^v$ , then (3.6) and (5.11) give

495 (5.13)  $\begin{aligned} [\text{Proj}_\Omega(x_k - \beta_k J_k^T c_k) - x_k]_i &= [v_k(\beta_k)]_i \\ &= [-\beta_k J_k^T c_k]_i - [-\beta_k J_k^T c_k - v_k(\beta_k)]_i = [-\beta_k J_k^T c_k]_i - [\beta_k w_k]_i, \end{aligned}$

496 so that (5.12) holds in this case. On the other hand, if  $i \in [n] \setminus \mathcal{A}_k^v$ , then  $[\text{Proj}_\Omega(x_k -$   
497  $\beta_k J_k^T c_k)]_i = [x_k + v_k(\beta_k)]_i = [x_k + v_k^c]_i > 0$  and  $[w_k]_i = 0$ . It follows that

498 (5.14)  $0 < [\text{Proj}_\Omega(x_k - \beta_k J_k^T c_k)]_i = \max \{ [x_k - \beta_k J_k^T c_k]_i, 0 \},$

499 which implies that  $[x_k - \beta_k J_k^T c_k]_i > 0$ . Combining this with  $[w_k]_i = 0$  shows that

500 (5.15) 
$$\begin{aligned} [\text{Proj}_\Omega(x_k - \beta_k J_k^T c_k) - x_k]_i &= [(x_k - \beta_k J_k^T c_k) - x_k]_i \\ &= [-\beta_k J_k^T c_k]_i = [-\beta_k J_k^T c_k - \beta_k w_k]_i \end{aligned}$$

501 so that (5.12) again holds for this case. This establishes that (5.12) holds, as claimed.  
502 It follows from the definition of  $v_k(\beta_k)$ , (5.12), and Assumption 5.1 that

503 
$$\begin{aligned} \left\| \frac{v_k(\beta_k)}{\beta_k} \right\|_2 &= \left\| \frac{\text{Proj}_\Omega(x_k - \beta_k J_k^T c_k) - x_k}{\beta_k} \right\|_2 = \left\| \frac{-\beta_k J_k^T c_k - \beta_k w_k}{\beta_k} \right\|_2 \\ &= \|J_k^T c_k + w_k\|_2 = \left\| \begin{bmatrix} J_k^T & I_{\mathcal{A}_k^v} \end{bmatrix} \begin{bmatrix} c_k \\ [w_k]_{\mathcal{A}_k^v} \end{bmatrix} \right\|_2 \\ &\geq \sigma_{\min}([J_k^T, I_{\mathcal{A}_k^v}]^T) \left\| \begin{bmatrix} c_k \\ [w_k]_{\mathcal{A}_k^v} \end{bmatrix} \right\|_2 \geq \sigma_{\min}\|c_k\|_2 \text{ for all } k \in \mathbb{N}, \end{aligned}$$

504 which completes the proof.  $\square$

505 We now give a bound on the improvement in linearized infeasibility at  $x_k$ .

506 LEMMA 5.10. *For all  $k \in \mathbb{N}$ , it holds that*

507 
$$\|c_k\|_2 - \|c_k + J_k s_k\|_2 = \|c_k\|_2 - \|c_k + J_k v_k\|_2 \geq \kappa_1 \sigma_{\min}^2 \|c_k\|_2 \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\}.$$

and

$$\|c_k\|_2 - \|c_k + J_k s_k\|_2 = \|c_k\|_2 - \|c_k + J_k v_k\|_2 \geq \frac{\kappa_1}{\kappa_c} \sigma_{\min}^2 \|c_k\|_2^2 \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\}.$$

508 *Proof.* It follows from (3.4) and Lemma 4.3 that  $\|c_k + J_k v_k\|_2 \leq \|c_k\|_2$ . It follows  
509 from this inequality and a difference-of-squares computation that

510 (5.16) 
$$\begin{aligned} \|c_k\|_2^2 - \|c_k + J_k v_k\|_2^2 &= (\|c_k\|_2 + \|c_k + J_k v_k\|_2)(\|c_k\|_2 - \|c_k + J_k v_k\|_2) \\ &\leq 2\|c_k\|_2(\|c_k\|_2 - \|c_k + J_k v_k\|_2). \end{aligned}$$

511 Combining (5.16), the third condition in (3.4), Lemma 4.4, and Lemma 5.9 we have

512 
$$2\|c_k\|_2(\|c_k\|_2 - \|c_k + J_k v_k\|_2) \geq \|c_k\|_2^2 - \|c_k + J_k v_k\|_2^2 = 2(m_k(0) - m_k(v_k))$$

513 
$$\geq 2(m_k(0) - m_k(v_k^c)) \geq 2\kappa_1 \left[ \frac{\|v_k(\beta_k)\|_2}{\beta_k} \right]^2 \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\}$$

514 
$$\geq 2\kappa_1 \sigma_{\min}^2 \|c_k\|_2^2 \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\}.$$

515 The proof of the first inequality follows by dividing through the previous inequality  
516 by  $2\|c_k\|_2$  and using the fact that  $J_k u_k = 0$  (see (4.8b)). The second inequality follows  
517 from the first inequality and the fact that  $\|c_k\|_2/\kappa_c \leq 1$  because of (3.12).  $\square$

518 We now establish that the merit parameter sequence is bounded away from zero.

519 LEMMA 5.11. *For all  $k \in \mathbb{N}$ , it holds that*

520 (5.17) 
$$\tau_{k,trial} \geq \tau_{\min,trial} := \frac{(1 - \sigma_c) \kappa_1 \sigma_{\min}^2 \min \left\{ \frac{1}{(1 + \kappa_J^2) \alpha_0}, \kappa_v \right\}}{2(\kappa_{\nabla f} + \kappa_{\partial r}) \kappa_v \kappa_J + 2\kappa_v^2 \kappa_J^2 \kappa_c} > 0 \text{ and}$$

521 (5.18) 
$$\tau_k \geq \tau_{\min} := \min\{\tau_0, (1 - \epsilon_\tau) \tau_{\min,trial}\} > 0.$$

522 *Proof.* We first prove (5.17). If  $A_k \leq 0$  in the definition of  $\tau_{k,\text{trial}}$ , then  $\tau_{k,\text{trial}} = \infty$   
523 so that (5.17) trivially holds. If  $A_k > 0$ , then it follows from the definition of  $\tau_{k,\text{trial}}$ ,  
524  $s_k = v_k + u_k$ ,  $J_k u_k = 0$  (see (4.8b)), Lemma 5.10, Lemma 5.4, the fact that  $\alpha_k \leq \alpha_0$   
525 for all  $k$  by construction of Algorithm 3.1, and (3.12) that

$$\begin{aligned}
526 \quad \tau_{k,\text{trial}} &= \frac{(1 - \sigma_c)(\|c_k\|_2 - \|c_k + J_k v_k\|_2)}{g_k^T s_k + \frac{1}{2\alpha_k} \|s_k\|_2^2 + r(x_k + s_k) - r_k} \\
527 \quad &\geq \frac{(1 - \sigma_c)\kappa_1 \sigma_{\min}^2 \|c_k\|_2 \min\left\{\frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k\right\}}{2(\kappa_{\nabla f} + \kappa_{\partial r})\kappa_v \kappa_J \alpha_k \|c_k\|_2 + 2\kappa_v^2 \kappa_J^2 \kappa_c \alpha_k \|c_k\|_2} \\
528 \quad &= \frac{(1 - \sigma_c)\kappa_1 \sigma_{\min}^2 \min\left\{\frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k\right\}}{2(\kappa_{\nabla f} + \kappa_{\partial r})\kappa_v \kappa_J \alpha_k + 2\kappa_v^2 \kappa_J^2 \kappa_c \alpha_k} \\
529 \quad &\geq \frac{(1 - \sigma_c)\kappa_1 \sigma_{\min}^2 \min\left\{\frac{1}{(1 + \kappa_J^2)\alpha_0}, \kappa_v\right\}}{2(\kappa_{\nabla f} + \kappa_{\partial r})\kappa_v \kappa_J + 2\kappa_v^2 \kappa_J^2 \kappa_c},
\end{aligned}$$

530 which proves (5.17). The merit parameter update rule (3.10) and (5.17) give (5.18).  $\square$

531 We may now state our worst-case complexity result for Algorithm 3.1.

532 **THEOREM 5.12.** *Suppose that Assumption 3.1 and Assumption 5.1 hold. Let  $\epsilon \in$   
533  $\mathbb{R}_{>0}$  be given. If  $\{k_1, k_2\} \subset \mathbb{N}$  are two iterations with  $k_1 < k_2$  such that  $k \in \mathcal{S}$  and  
534  $\chi_k > \epsilon$  for all iterations  $k_1 \leq k < k_2$ , then it follows that*

$$535 \quad (5.19) \quad k_2 - k_1 \leq \left\lfloor \frac{\tau_0(f(x_0) + r(x_0) - f_{\inf}) + \|c(x_0)\|_2}{\kappa_{\Phi} \epsilon^2} \right\rfloor$$

536 with  $\kappa_{\Phi} = \eta_{\Phi} \min\left\{\frac{\tau_{\min} \alpha_{\min}}{8}, \frac{\tau_{\min}}{8\alpha_0}, \frac{\sigma_c \kappa_1}{\kappa_c} \sigma_{\min}^2 \min\left\{\frac{1}{1 + \kappa_J^2}, \kappa_v \alpha_{\min}\right\}\right\}$ . Moreover, the max-  
537 imum number of iterations before  $\chi_k \leq \epsilon$  for some iteration  $k \in \mathbb{N}$  is

$$538 \quad \left( \max \left\{ 0, \left\lceil \frac{\log\left(\frac{\tau_{\min}}{2\alpha_0(\tau_{\min} L_g + L_J)}\right)}{\log(\xi)} \right\rceil \right\} + 1 \right) \left\lfloor \frac{\tau_0(f(x_0) - f_{\inf} + r(x_0)) + \|c(x_0)\|_2}{\kappa_{\Phi} \epsilon^2} \right\rfloor.$$

539 *Proof.* Let  $\{k_1, k_2\}$  be as in the statement of the theorem. Then, for all  $k \in \mathcal{S}$  and  
540  $k_1 \leq k < k_2$ , it follows from Lemma 5.7(i)–(ii),  $k \in \mathcal{S}$ , (3.12), the second inequality  
541 of Lemma 5.10, and Lemma 5.6 that

$$\begin{aligned}
542 \quad (5.20) \quad &\bar{\Phi}_{\tau_k}(x_k) - \bar{\Phi}_{\tau_{k+1}}(x_{k+1}) \geq \bar{\Phi}_{\tau_k}(x_k) - \bar{\Phi}_{\tau_k}(x_{k+1}) = \Phi_{\tau_k}(x_k) - \Phi_{\tau_k}(x_{k+1}) \\
&\geq \eta_{\Phi} \left( \frac{\tau_k}{4\alpha_k} \|s_k\|_2^2 + \sigma_c (\|c_k\|_2 - \|c_k + J_k s_k\|_2) \right) \\
&\geq \eta_{\Phi} \left[ \frac{\tau_k \alpha_k}{4} \left( \frac{\|s_k\|_2}{\alpha_k} \right)^2 + \sigma_c \left( \frac{\kappa_1}{\kappa_c} \sigma_{\min}^2 \|c_k\|_2^2 \min\left\{\frac{1}{1 + \kappa_J^2}, \kappa_v \alpha_k\right\} \right) \right] \\
&= \eta_{\Phi} \left[ \frac{\tau_k \alpha_k}{8} \left( \frac{\|s_k\|_2}{\alpha_k} \right)^2 + \frac{\tau_k \|s_k\|_2^2}{8\alpha_k} + \sigma_c \left( \frac{\kappa_1}{\kappa_c} \sigma_{\min}^2 \|c_k\|_2^2 \min\left\{\frac{1}{1 + \kappa_J^2}, \kappa_v \alpha_{\min}\right\} \right) \right].
\end{aligned}$$

543 Lemma 4.6, (5.20), (4.8), (5.18), (5.4), and  $\alpha_k \leq \alpha_0$  for all  $k \geq 0$  give

$$\begin{aligned}
544 \quad &\bar{\Phi}_{\tau_k}(x_k) - \bar{\Phi}_{\tau_{k+1}}(x_{k+1}) \\
545 \quad &\geq \eta_{\Phi} \left[ \frac{\tau_k \alpha_k}{8} \|g_k + g_{r,k} + J_k^T y_k + z_k\|_2^2 + \frac{\tau_k}{8\alpha_k} \|\min\{x_k, -z_k\}\|_2^2 \right]
\end{aligned}$$

$$\begin{aligned}
& + \sigma_c \left( \frac{\kappa_1}{\kappa_c} \sigma_{\min}^2 \|c_k\|_2^2 \min \left\{ \frac{1}{1+\kappa_J^2}, \kappa_v \alpha_{\min} \right\} \right) \Big] \\
& \geq \eta_\Phi \left[ \frac{\tau_{\min} \alpha_{\min}}{8} \|g_k + g_{r,k} + J_k^T y_k + z_k\|_2^2 + \frac{\tau_{\min}}{8\alpha_0} \|\min\{x_k, -z_k\}\|_2^2 \right. \\
& \quad \left. + \sigma_c \left( \frac{\kappa_1}{\kappa_c} \sigma_{\min}^2 \|c_k\|_2^2 \min \left\{ \frac{1}{1+\kappa_J^2}, \kappa_v \alpha_{\min} \right\} \right) \right] \\
& \geq \kappa_\Phi \chi_k^2
\end{aligned}$$

550 where  $\kappa_\Phi$  is defined in the statement of the current theorem. Using this inequality,  
551 Lemma 5.7(iii), and nonnegativity of  $\bar{\Phi}_\tau$  for all  $\tau \in \mathbb{R}_{>0}$ , we find that

$$\begin{aligned}
552 \quad & \bar{\Phi}_{\tau_0}(x_0) \geq \bar{\Phi}_{\tau_{k_1}}(x_{k_1}) \geq \bar{\Phi}_{\tau_{k_1}}(x_{k_1}) - \bar{\Phi}_{\tau_{k_2}}(x_{k_2}) \\
553 \quad & = \sum_{k=k_1}^{k_2-1} (\bar{\Phi}_{\tau_k}(x_k) - \bar{\Phi}_{\tau_{k+1}}(x_{k+1})) \geq \sum_{k=k_1}^{k_2-1} \kappa_\Phi \chi_k^2,
\end{aligned}$$

which may be combined with  $\chi_k > \epsilon$  for all iterations  $k_1 \leq k \leq k_2$  to conclude that

$$\bar{\Phi}_{\tau_0}(x_0) \geq (k_2 - k_1) \kappa_\Phi \epsilon^2,$$

554 from which (5.8) follows. The final result in the theorem, namely the claimed upper  
555 bound on the maximum iterations before  $\chi_k \leq \epsilon$ , follows from what we just proved  
556 and the fact that maximum number of unsuccessful iterations is bounded as in (5.5).  $\square$

557 **5.2.3. Analysis under a limit-point constraint qualification.** The analysis  
558 in this section is performed under Assumption 3.1 and the following two assumptions.  
559 Before stating them, we remark that all of the results from Section 5.2.1 still hold.

560 **ASSUMPTION 5.2.** *The set  $\mathcal{X}$  in Assumption 3.1 is bounded.*

561 **ASSUMPTION 5.3.** *Let  $\mathcal{L}$  denote the set of limit points of the sequence  $\{x_k\}$  generated  
562 by Algorithm 3.1. Every  $x_* \in \mathcal{L}$  satisfies the LICQ, i.e., if  $x_* \in \mathcal{L}$ , then  
563  $[J(x_*)^T, I_{\mathcal{A}(x_*)}^T]^T$  has full row rank with  $I_{\mathcal{A}(x_*)}$  denoting the subset of the rows of the  
564 identity matrix  $I$  that corresponds to the index set  $\mathcal{A}(x_*) := \{i \in [n] : [x_*]_i = 0\}$ .*

565 The previous assumption has important consequences in terms of a certain type  
566 of infeasible point (see Lemma 4.5(ii)), as we now define.

567 **DEFINITION 5.13.** *We say that  $\bar{x} \in \mathbb{R}^n$  is an infeasible stationary point (ISP) for  
568 problem (1.1) if and only if  $\bar{x} \in \Omega$ ,  $\bar{x} = \text{Proj}_\Omega(\bar{x} - J(\bar{x})^T c(\bar{x}))$ , and  $c(\bar{x}) \neq 0$ .*

569 We now show that any limit point of the sequence of iterates cannot be an ISP.

570 **LEMMA 5.14.** *If  $x_*$  is a limit point of  $\{x_k\}$ , then  $x_*$  cannot be an ISP.*

571 *Proof.* Let  $x_* \in \mathbb{R}^n$  be a limit point of  $\{x_k\}$ . Suppose that  $x_* \in \Omega$  and  $x_* =$   
572  $\text{Proj}_\Omega(x_* - J(x_*)^T c(x_*))$ . The proof will be complete if we can show that  $c(x_*) = 0$   
573 since this would prove that  $x_*$  is not an ISP. Thus, we now prove that  $c(x_*) = 0$ .

574 It follows using the same proof as in Lemma 4.5 with  $x_k$  replaced by  $x_*$  that  
575  $x_* = \text{Proj}_\Omega(x_* - J(x_*)^T c(x_*))$  implies that  $x_*$  is a first-order KKT point for the  
576 feasibility problem (3.3). Therefore, there exists  $z_* \in \mathbb{R}_{\geq 0}^n$  satisfying  $x_* \cdot z_* = 0$   
577 (componentwise), and  $J(x_*)^T c(x_*) = z_*$ . It follows from these equations and  $\mathcal{I}(x_*) =$   
578  $[n] \setminus \mathcal{A}(x_*)$  that  $[J(x_*)^T c(x_*)]_{\mathcal{I}(x_*)} = 0$ , where we also note that  $\mathcal{I}(x_*) \neq \emptyset$  as a  
579 consequence of Assumption 5.3. Letting  $J_{\mathcal{I}(x_*)}(x_*)$  denote the columns of  $J(x_*)$  that  
580 correspond to the indices in  $\mathcal{I}(x_*)$ , it follows from above that  $0 = [J(x_*)^T c(x_*)]_{\mathcal{I}(x_*)} =$   
581  $[J_{\mathcal{I}(x_*)}(x_*)^T c(x_*)]$ . Since  $J_{\mathcal{I}(x_*)}(x_*)$  must have full row rank (see [44, Lemma 2.1.3]),  
582 it follows that  $c(x_*) = 0$ , which completes the proof.  $\square$

583 The next result bounds  $\|v_k(1)\|_2$  by the infeasibility of the equality constraints.

584 LEMMA 5.15. *For all  $k \in \mathbb{N}$ , it holds that  $\|v_k(1)\|_2 \leq \kappa_J \|c_k\|_2$ .*

585 *Proof.* It follows from the definition of  $v_k(1)$  in (3.6),  $x_k \in \Omega$  for all  $k \in \mathbb{N}$  by how

586 Algorithm 3.1 is designed, non-expansivity of the projection operator, and (3.12) that

$$587 \|v_k(1)\|_2 = \|\text{Proj}_\Omega(x_k - J_k^T c_k) - x_k\|_2 = \|\text{Proj}_\Omega(x_k - J_k^T c_k) - \text{Proj}_\Omega(x_k)\|_2 \\ 588 \leq \|J_k^T c_k\|_2 \leq \kappa_J \|c_k\|_2,$$

589 which completes the proof.  $\square$

590 We can now prove that our infeasiblity measure converges to zero.

591 LEMMA 5.16. *The iterate sequence  $\{x_k\}$  satisfies  $\lim_{k \rightarrow \infty} \|v_k(1)\|_2 = 0$ .*

592 *Proof.* From Theorem 5.12, it follows that there exists a subsequence  $\mathcal{K}_1 \subseteq \mathbb{N}$  such  
593 that  $\lim_{k \in \mathcal{K}_1} \|v_k(1)\|_2 = 0$ . Now, for the purpose of reaching a contradiction, assume  
594 that there exists a subsequence of iterations  $\mathcal{K}_2 \subseteq \mathbb{N} \setminus \mathcal{K}_1$  and a scalar  $v_{\min} \in \mathbb{R}_{>0}$   
595 such that  $\|v_k(1)\|_2 \geq v_{\min}$  for all  $k \in \mathcal{K}_2$ . We now proceed by considering two cases.

596 **Case 1:**  $\{\tau_k\} \rightarrow 0$ . The definitions of  $\mathcal{K}_1$  and  $\mathcal{K}_2$  allow us to define, for each  $k \in \mathcal{K}_1$ ,  
597 the quantity  $\hat{k}(k)$  as the smallest iteration in  $\mathcal{K}_2$  that is strictly larger than  $k$ . We can  
598 use this definition, Lemma 5.15,  $\{\tau_k\} \rightarrow 0$ , (3.12), Lemma 5.7(iii), and nonnegativity  
599 of  $r$  to conclude that the following holds for each sufficiently large  $k \in \mathcal{K}_1$ :

$$600 \frac{v_{\min}}{2\kappa_J} \leq \frac{\|c(x_{\hat{k}(k)})\|_2}{2} \leq \tau_{\hat{k}(k)}(f_{\hat{k}(k)} - f_{\inf} + r(x_{\hat{k}(k)})) + \|c(x_{\hat{k}(k)})\|_2 = \bar{\Phi}_{\tau_{\hat{k}(k)}}(x_{\hat{k}(k)}) \\ 601 \leq \bar{\Phi}_{\tau_k}(x_k) = \tau_k(f_k - f_{\inf} + r(x_k)) + \|c(x_k)\|_2 \leq 2\|c_k\|_2.$$

It follows from this inequality and the definition of  $\mathcal{K}_1$  that

$$\lim_{k \in \mathcal{K}_1} \|v_k(1)\|_2 = 0 \text{ and } \liminf_{k \in \mathcal{K}_1} \|c_k\|_2 \geq \frac{v_{\min}}{2\kappa_J} > 0.$$

602 Therefore, every limit point of  $\{x_k\}_{k \in \mathcal{K}_1}$  must be an ISP, and at least one such limit  
603 point must exist as a consequence of Assumption 5.2. This contradicts Lemma 5.14.

604 **Case 2:  $\{\tau_k\}$  is bounded away from zero.** In this case, it follows from The-  
605 orem 5.8(i) that the proximal parameter sequence  $\{\alpha_k\}$  is also bounded away from  
606 zero. Given the manner in which both sequences are defined in Algorithm 3.1, we can  
607 conclude that there exists  $\hat{k} \in \mathbb{N}$  such that  $\tau_k = \tau_{\hat{k}} > 0$  and  $\alpha_k = \alpha_{\hat{k}} > 0$  for all  $k \geq \hat{k}$ .  
608 We may now use the same logic as in the proof of Lemma 5.8(i) and (3.12) to obtain

$$609 \infty > \bar{\Phi}_{\tau_0}(x_0) \geq \sum_{k=0}^{\infty} (\bar{\Phi}_{\tau_k}(x_k) - \bar{\Phi}_{\tau_{k+1}}(x_{k+1})) \\ \geq \sum_{\hat{k} \leq k \in \mathcal{S}} (\bar{\Phi}_{\tau_k}(x_k) - \bar{\Phi}_{\tau_{k+1}}(x_{k+1})) \\ \geq \sum_{\hat{k} \leq k \in \mathcal{S}} \eta_{\Phi} \frac{\sigma_c \kappa_1}{\kappa_c} \alpha_{\hat{k}} \|v_k(1)\|_2^2 \min \left\{ \frac{1}{1+\kappa_J^2}, \kappa_v \alpha_{\hat{k}} \right\},$$

610 which implies that  $\lim_{k \in \mathcal{S}} \|v_k(1)\|_2 = 0$ . Combining this result with the fact that  
611  $x_{k+1} = x_k$  whenever  $k \notin \mathcal{S}$  and that the definition of  $v_k(1)$  depends only on  $x_k$ , the  
612 projection onto  $\Omega$  (which is continuous), and the continuous functions  $c$  and  $J$ , it  
613 follows that  $\lim_{k \rightarrow \infty} \|v_k(1)\|_2 = 0$ . This contradicts the definition of  $\mathcal{K}_2$ .

614 Since we have shown that both **Case 1** and **Case 2** cannot occur, and these are  
615 the only cases that can possibly occur, we must conclude that our original assumption  
616 was incorrect, namely the existence of the set  $\mathcal{K}_2$ . This completes the proof.  $\square$

617 Next, we formally establish that  $\mathcal{L}$  is a compact set.

618 **LEMMA 5.17.** *The set  $\mathcal{L}$  in Assumption 5.3 is compact.*

*Proof.* By Assumption 5.2, the set  $\mathcal{L}$  is bounded. It remains to show that  $\mathcal{L}$  is closed. To this end, suppose that  $\{x_j^{\mathcal{L}}\}_{j \geq 1} \subseteq \mathcal{L}$  and  $x^{\mathcal{L}} \in \mathbb{R}^n$  satisfy  $\lim_{j \rightarrow \infty} x_j^{\mathcal{L}} = x^{\mathcal{L}}$ ; we prove that  $x^{\mathcal{L}} \in \mathcal{L}$ . Let us define a sequence  $\mathcal{K} = \{k_1, k_2, \dots\} \subseteq \mathbb{N}$ . In particular, let  $k_1$  be the smallest integer such that the iterate  $x_{k_1}$  satisfies  $\|x_1^{\mathcal{L}} - x_{k_1}\|_2 \leq 1$ . We then iteratively define  $k_j$  for  $j \geq 2$  as the smallest integer  $k_j$  such that  $k_j > k_{j-1}$  and the iterate  $x_{k_j}$  satisfies  $\|x_j^{\mathcal{L}} - x_{k_j}\|_2 \leq 1/j$ . In summary,  $\mathcal{K} = \{k_1, k_2, \dots\} \subseteq \mathbb{N}$  is a strictly monotonically increasing subsequence of  $\mathbb{N}$  such that  $\|x_j^{\mathcal{L}} - x_{k_j}\|_2 \leq 1/j$  for all  $j$ . It follows from this inequality and the triangle inequality that

$$\|x^{\mathcal{L}} - x_{k_j}\|_2 \leq \|x^{\mathcal{L}} - x_j^{\mathcal{L}}\|_2 + \|x_j^{\mathcal{L}} - x_{k_j}\|_2 \leq \|x^{\mathcal{L}} - x_j^{\mathcal{L}}\|_2 + \frac{1}{j} \text{ for all } j \geq 1.$$

619 Combining this inequality with  $\lim_{j \rightarrow \infty} x_j^{\mathcal{L}} = x^{\mathcal{L}}$ , it follows that  $\lim_{j \rightarrow \infty} x_{k_j} = x^{\mathcal{L}}$ ,  
620 which proves that  $x^{\mathcal{L}} \in \mathcal{L}$  as claimed, thus completing the proof.  $\square$

621 The next key lemma uses the function  $\delta(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}_{>0}$  defined as

$$622 \quad (5.21) \quad \delta_{\min}(x) := \min_{i \in \mathcal{I}(x)} [x]_i,$$

623 which gives a measure for how far the inactive variables at  $x$  are from being active.

624 **LEMMA 5.18.** *The following hold for the set of limit points  $\mathcal{L}$ :*

625 (i) *There exist  $n_{\mathcal{L}} \in \mathbb{N}$ ,  $\{x_i^{\mathcal{L}}\}_{i=1}^{n_{\mathcal{L}}} \subseteq \mathcal{L}$ , and  $\{\epsilon_i^{\mathcal{L}}\}_{i=1}^{n_{\mathcal{L}}} \subset \mathbb{R}_{>0}$  such that*  
626 (a)  *$\mathcal{L} \subset \bigcup_{i=1}^{n_{\mathcal{L}}} \mathcal{B}(x_i^{\mathcal{L}}, \epsilon_i^{\mathcal{L}})$ , and*  
627 (b) *if, for some  $j$ , it holds that  $x \in \mathcal{B}(x_j^{\mathcal{L}}, \epsilon_j^{\mathcal{L}})$ , then*

$$628 \quad (5.22a) \quad \|x - x_j^{\mathcal{L}}\|_2 \leq \frac{1}{3} \delta_{\min}(x_j^{\mathcal{L}}),$$

$$629 \quad (5.22b) \quad \mathcal{A}(x) \subseteq \mathcal{A}(x_j^{\mathcal{L}}), \text{ and}$$

$$630 \quad (5.22c) \quad \sigma_{\min}([J(x)^T, I_{\mathcal{A}(x_j^{\mathcal{L}})}^T]^T) \geq \frac{1}{2} \sigma_{\min}([J(x_j^{\mathcal{L}})^T, I_{\mathcal{A}(x_j^{\mathcal{L}})}^T]^T).$$

631 (ii) *For the objects in part (i), there exists  $\epsilon_{\min}^{\mathcal{L}} \in \mathbb{R}_{>0}$  such that if  $\bar{x} \in \mathbb{R}^n$  satisfies  
632  $\text{dist}(\bar{x}, \mathcal{L}) \leq \epsilon_{\min}^{\mathcal{L}}$ , then  $\bar{x} \in \bigcup_{i=1}^{n_{\mathcal{L}}} \mathcal{B}(x_i^{\mathcal{L}}, \epsilon_i^{\mathcal{L}})$  and there exists  $j \in [n_{\mathcal{L}}]$  such that*

$$633 \quad \sigma_{\min}([J(\bar{x})^T, I_{\mathcal{A}(x_j^{\mathcal{L}})}^T]^T) \geq \min_{i \in [n_{\mathcal{L}}]} \frac{1}{2} \sigma_{\min}([J(x_i^{\mathcal{L}})^T, I_{\mathcal{A}(x_i^{\mathcal{L}})}^T]^T) =: \sigma_{\min}^{\mathcal{L}} > 0,$$

634 where the inequality  $\sigma_{\min}^{\mathcal{L}} > 0$  is a consequence of Assumption 5.3.

635 *Proof.* For  $x^{\mathcal{L}} \in \mathcal{L}$ , let  $\epsilon(x^{\mathcal{L}}) \in \mathbb{R}_{>0}$  satisfy that if  $x \in \mathcal{B}(x^{\mathcal{L}}, \epsilon(x^{\mathcal{L}}))$  then  $\mathcal{I}(x^{\mathcal{L}}) \subseteq \mathcal{I}(x)$ ,  $\|x - x^{\mathcal{L}}\|_2 \leq \frac{1}{3} \delta_{\min}(x^{\mathcal{L}})$ , and  $\sigma_{\min}([J(x)^T, I_{\mathcal{A}(x^{\mathcal{L}})}^T]^T) \geq \frac{\sigma_{\min}}{2} ([J(x^{\mathcal{L}})^T, I_{\mathcal{A}(x^{\mathcal{L}})}^T]^T)$ ,  
636 where satisfying the third condition is possible because of the continuity of singular values of a matrix with respect to its entries and Assumption 5.3. It follows  
637 that  $\bigcup_{x^{\mathcal{L}} \in \mathcal{L}} \mathcal{B}(x^{\mathcal{L}}, \epsilon(x^{\mathcal{L}}))$  is an open cover of the compact set  $\mathcal{L}$  (see Lemma 5.17).  
638 Using this fact and the definition of a compact set, it follows that there exists a  
639 finite subcover, i.e., there exist  $n_{\mathcal{L}} \in \mathbb{N}$ ,  $\{x_i^{\mathcal{L}}\}_{i=1}^{n_{\mathcal{L}}} \subseteq \mathcal{L}$ , and  $\{\epsilon_i^{\mathcal{L}}\}_{i=1}^{n_{\mathcal{L}}} \subset \mathbb{R}_{>0}$  such  
640 that  $\mathcal{L} \subset \bigcup_{i=1}^{n_{\mathcal{L}}} \mathcal{B}(x_i^{\mathcal{L}}, \epsilon_i^{\mathcal{L}})$  and if, for some  $j \in \{1, 2, \dots, n_{\mathcal{L}}\}$ , it holds that  $x \in$   
642

643  $\mathcal{B}(x_j^{\mathcal{L}}, \epsilon_j^{\mathcal{L}})$  then  $\mathcal{I}(x_j^{\mathcal{L}}) \subseteq \mathcal{I}(x)$ ,  $\|x - x_j^{\mathcal{L}}\|_2 \leq \frac{1}{3}\delta_{\min}(x_j^{\mathcal{L}})$ , and  $\sigma_{\min}([J(x)^T, I_{\mathcal{A}(x_j^{\mathcal{L}})}^T]^T) \geq$   
644  $\frac{1}{2}\sigma_{\min}([J(x_j^{\mathcal{L}})^T, I_{\mathcal{A}(x_j^{\mathcal{L}})}^T]^T)$ . Since  $\mathcal{I}(x_j^{\mathcal{L}}) \subseteq \mathcal{I}(x)$  is equivalent to  $\mathcal{A}(x) \subseteq \mathcal{A}(x_j^{\mathcal{L}})$ , we  
645 have completed the proof of part (i).

646 We now prove part (ii). First, using the *finite* subcover computed in part (i)  
647 and the fact that  $\mathcal{L}$  is compact, there exists  $\epsilon_{\min}^{\mathcal{L}} \in \mathbb{R}_{>0}$  such that if  $x \in \mathbb{R}^n$  sat-  
648 isfies  $\text{dist}(x, \mathcal{L}) \leq \epsilon_{\min}^{\mathcal{L}}$ , then  $x \in \cup_{i=1}^{n_{\mathcal{L}}} \mathcal{B}(x_i^{\mathcal{L}}, \epsilon_i^{\mathcal{L}})$ . Let  $\bar{x}$  be an arbitrary point that  
649 satisfies  $\text{dist}(\bar{x}, \mathcal{L}) \leq \epsilon_{\min}^{\mathcal{L}}$ . Then, it follows that there exists  $j \in \{1, 2, \dots, n_{\mathcal{L}}\}$   
650 such that  $\bar{x} \in \mathcal{B}(x_j^{\mathcal{L}}, \epsilon_j^{\mathcal{L}})$ , which combined with part (i)(b) gives  $\mathcal{A}(\bar{x}) \subseteq \mathcal{A}(x_j^{\mathcal{L}})$  and  
651  $\sigma_{\min}([J(\bar{x})^T, I_{\mathcal{A}(x_j^{\mathcal{L}})}^T]^T) \geq \frac{1}{2}\sigma_{\min}([J(x_j^{\mathcal{L}})^T, I_{\mathcal{A}(x_j^{\mathcal{L}})}^T]^T) \geq \sigma_{\min}^{\mathcal{L}} > 0$ , as claimed.  $\square$

652 The next result shows that iterates of the algorithm eventually satisfy the prop-  
653 erties of the previous lemma.

654 LEMMA 5.19. *There exists  $\bar{k} \in \mathbb{N}$  such that, for each  $k \geq \bar{k}$ , there exists a corre-  
655 sponding  $j \in [n_{\mathcal{L}}]$  that satisfies, with  $\sigma_{\min}^{\mathcal{L}}$  defined in Lemma 5.18(ii), the following:*  
656

$$657 \quad (5.23a) \quad \|x_k - x_j^{\mathcal{L}}\|_2 \leq \frac{1}{3}\delta_{\min}(x_j^{\mathcal{L}}),$$

$$658 \quad (5.23b) \quad \mathcal{A}(x_k) \subseteq \mathcal{A}(x_j^{\mathcal{L}}), \text{ and}$$

$$659 \quad (5.23c) \quad \sigma_{\min}([J_k^T, I_{\mathcal{A}(x_j^{\mathcal{L}})}^T]^T) \geq \frac{1}{2}\sigma_{\min}([J(x_j^{\mathcal{L}})^T, I_{\mathcal{A}(x_j^{\mathcal{L}})}^T]^T) \geq \sigma_{\min}^{\mathcal{L}} > 0.$$

660 *Proof.* Let  $\epsilon_{\min}^{\mathcal{L}} > 0$  be defined as in Lemma 5.18(ii). Since  $\mathcal{L}$  is the set of all limit  
661 points, there exists an iteration  $\bar{k}$  such that  $\text{dist}(x_k, \mathcal{L}) \leq \epsilon_{\min}^{\mathcal{L}}$  for all  $k \geq \bar{k}$  (this  $\bar{k}$   
662 is now the  $\bar{k}$  whose existence is claimed in the statement of the current lemma). For  
663 the remainder of the proof, consider arbitrary  $k \geq \bar{k}$ . It follows from the definition  
664 of  $\bar{k}$  that  $\text{dist}(x_k, \mathcal{L}) \leq \epsilon_{\min}^{\mathcal{L}}$ , and then from Lemma 5.18(ii) that there exists  $j \in [n_{\mathcal{L}}]$   
665 such that  $x_k \in \mathcal{B}(x_j^{\mathcal{L}}, \epsilon_j^{\mathcal{L}})$ . Conditions (5.23a)–(5.23c) now follow from Lemma 5.18.  $\square$

666 We now give a lower bound on  $\|v_k(1)\|_2$  in terms of  $\|c_k\|_2$ , which is crucial to  
667 giving a lower bound on the merit parameter sequence. The result uses the constant

$$668 \quad (5.24) \quad \delta_{\min}^{\mathcal{L}} := \min_{j \in [n_{\mathcal{L}}]} \delta_{\min}(x_j^{\mathcal{L}}) > 0.$$

669 LEMMA 5.20. *For all sufficiently large  $k \in \mathbb{N}$ , it holds that  $\|v_k(1)\|_2 \geq \sigma_{\min}^{\mathcal{L}} \|c_k\|_2$ ,  
670 where the positive constant  $\sigma_{\min}^{\mathcal{L}}$  is defined in Lemma 5.18(ii).*

671 *Proof.* With  $\delta_{\min}^{\mathcal{L}}$  in (5.24), Lemma 5.16 ensures the existence  $\bar{k}_1$  such that

$$672 \quad (5.25) \quad \|v_k(1)\|_2 = \|\text{Proj}_{\Omega}(x_k - J_k^T c_k) - x_k\|_2 \leq \frac{1}{3}\delta_{\min}^{\mathcal{L}} \quad \text{for all } k \geq \bar{k}_1.$$

673 Let  $\{\epsilon_{\min}^{\mathcal{L}}, \sigma_{\min}^{\mathcal{L}}\} \subset \mathbb{R}_{>0}$  be as stated in Lemma 5.18, and let  $\bar{k}_2$  play the role of  $\bar{k}$  from  
674 Lemma 5.19. For the remainder of the proof, consider arbitrary  $k \geq \max\{\bar{k}_1, \bar{k}_2\}$ . It  
675 follows from the definition of  $\bar{k}_2$  that  $x_k$  satisfies (5.23a)–(5.23c) for some  $j \in [n_{\mathcal{L}}]$ .  
676 Using (5.25), (5.23a), and definitions of  $\delta_{\min}(x_j^{\mathcal{L}})$  and  $\delta_{\min}^{\mathcal{L}}$ , each  $i \in \mathcal{I}(x_j^{\mathcal{L}})$  satisfies

$$\begin{aligned} 677 \quad [\text{Proj}_{\Omega}(x_k - J_k^T c_k)]_i &\geq [x_k]_i - \frac{1}{3}\delta_{\min}^{\mathcal{L}} \geq [x_j^{\mathcal{L}}]_i - \frac{1}{3}\delta_{\min}(x_j^{\mathcal{L}}) - \frac{1}{3}\delta_{\min}^{\mathcal{L}} \\ &\geq \delta_{\min}(x_j^{\mathcal{L}}) - \frac{1}{3}\delta_{\min}(x_j^{\mathcal{L}}) - \frac{1}{3}\delta_{\min}^{\mathcal{L}} = \frac{2}{3}\delta_{\min}(x_j^{\mathcal{L}}) - \frac{1}{3}\delta_{\min}^{\mathcal{L}} \\ &\geq \frac{2}{3}\delta_{\min}^{\mathcal{L}} - \frac{1}{3}\delta_{\min}^{\mathcal{L}} = \frac{1}{3}\delta_{\min}^{\mathcal{L}}. \end{aligned}$$

678 Hence, for all  $i \in \mathcal{I}(x_j^{\mathcal{L}})$  it holds that  $[x_k - J_k^T c_k]_i > 0$ . Now, define  $w_k \in \mathbb{R}^n$  as

$$679 \quad (5.26) \quad [w_k]_i = \begin{cases} 0 & \text{if } i \in \mathcal{I}(x_j^{\mathcal{L}}), \\ -[J_k^T c_k]_i - [v_k(1)]_i & \text{if } i \in \mathcal{A}(x_j^{\mathcal{L}}). \end{cases}$$

680 The definition of  $w_k$ , the fact that  $[x_k - J_k^T c_k]_i > 0$  for all  $i \in \mathcal{I}(x_j^{\mathcal{L}})$ , and (5.23c) give

$$681 \quad \begin{aligned} \|v_k(1)\|_2 &= \|\text{Proj}_{\Omega}(x_k - J_k^T c_k) - x_k\|_2 = \|-J_k^T c_k - w_k\|_2 \\ &= \left\| \begin{bmatrix} J_k^T, I_{\mathcal{A}(x_j^{\mathcal{L}})}^T \end{bmatrix} \begin{bmatrix} c_k \\ [w_k]_{\mathcal{A}(x_j^{\mathcal{L}})} \end{bmatrix} \right\|_2 \geq \sigma_{\min}^{\mathcal{L}} \|c_k\|_2, \end{aligned}$$

682 which completes the proof.  $\square$

683 Our next result gives a new bound on the model decrease.

684 LEMMA 5.21. *For  $\kappa_1 \in (0, 1]$  in Lemma 4.3, all sufficiently large  $k \in \mathbb{N}$  satisfy*

$$685 \quad (5.27) \quad \|c_k\|_2 - \|c_k + J_k v_k^c\|_2 \geq \kappa_1 (\sigma_{\min}^{\mathcal{L}})^2 \|c_k\|_2 \min \left\{ \frac{1}{1 + \kappa_J^2}, \kappa_v \alpha_k \right\}.$$

686 *Proof.* If  $\delta_k = 0$ , then either  $\|c_k\|_2 = 0$  and the inequality holds trivially, or  
687  $\|c_k\|_2 \neq 0$  and the algorithm terminates finitely, which is a contradiction to our overall  
688 setting in this subsection that the algorithm does not terminate finitely. Therefore, we  
689 may proceed assuming  $\delta_k \neq 0$ . It follows from Lemma 4.3 that  $\|c_k + J_k v_k^c\|_2 \leq \|c_k\|_2$ .  
690 Using this inequality and a difference-of-squares computation, we have that

$$691 \quad (5.28) \quad \begin{aligned} \|c_k\|_2^2 - \|c_k + J_k v_k^c\|_2^2 &= (\|c_k\|_2 + \|c_k + J_k v_k^c\|_2)(\|c_k\|_2 - \|c_k + J_k v_k^c\|_2) \\ &\leq 2\|c_k\|_2(\|c_k\|_2 - \|c_k + J_k v_k^c\|_2). \end{aligned}$$

692 Combining (5.28), (4.4), Lemma 5.20, and (3.12), all sufficiently large  $k \in \mathbb{N}$  satisfy

$$693 \quad \begin{aligned} 2\|c_k\|_2(\|c_k\|_2 - \|c_k + J_k v_k^c\|_2) &\geq \|c_k\|_2^2 - \|c_k + J_k v_k^c\|_2^2 = 2(m_k(0) - m_k(v_k^c)) \\ 694 &\geq 2\kappa_1 \|v_k(1)\|_2^2 \min \left\{ \frac{1}{1 + \|J_k^T J_k\|_2}, \kappa_v \alpha_k \right\} \\ 695 &\geq 2\kappa_1 (\sigma_{\min}^{\mathcal{L}})^2 \|c_k\|_2^2 \min \left\{ \frac{1}{1 + \kappa_J^2}, \kappa_v \alpha_k \right\}. \end{aligned}$$

696 If  $\|c_k\|_2 = 0$ , then again the desired inequality holds trivially. Otherwise, dividing the  
697 above inequality by  $2\|c_k\|_2$  gives (5.27), and thus completes the proof.  $\square$

698 We now bound the merit and proximal parameter sequences away from zero.

699 LEMMA 5.22. *Let  $\bar{k} > 0$  be sufficiently large that the results in Lemma 5.20 and  
700 Lemma 5.21 hold. Then, each  $k \geq \bar{k}$  yields*

$$701 \quad (5.29) \quad \tau_{k,\text{trial}} \geq \bar{\tau}_{\min,\text{trial}} := \frac{(1 - \sigma_c)\kappa_1 (\sigma_{\min}^{\mathcal{L}})^2 \min \left\{ \frac{1}{(1 + \kappa_J^2)\alpha_0}, \kappa_v \right\}}{2(\kappa_{\nabla f} + \kappa_{\partial r})\kappa_v \kappa_J + 2\kappa_v^2 \kappa_J^2 \kappa_c} > 0.$$

702 *The merit parameter sequence itself satisfies, for all  $k \in \mathbb{N}$ , the inequality*

$$703 \quad (5.30) \quad \tau_k \geq \bar{\tau}_{\min} := \min\{\tau_{\bar{k}-1}, (1 - \epsilon_{\tau})\bar{\tau}_{\min,\text{trial}}\} > 0.$$

704 *Finally, the proximal parameter sequence satisfies, for all  $k \in \mathbb{N}$ , the inequality*

$$705 \quad (5.31) \quad \alpha_k \geq \bar{\alpha}_{\min} := \min\{\alpha_0, \frac{\xi \bar{\tau}_{\min}}{2(\bar{\tau}_{\min} L_g + L_J)}\} > 0.$$

706 *Proof.* We first prove (5.29). If  $A_k \leq 0$  in the definition of  $\tau_{k,\text{trial}}$ , then  $\tau_{k,\text{trial}} = \infty$   
707 so that (5.29) trivially holds. If  $A_k > 0$ , then it follows from the definition of  $\tau_{k,\text{trial}}$ ,

708  $s_k = v_k + u_k$ ,  $J_k u_k = 0$  (see (4.8b)), Lemma 5.4, Lemma 5.21, the fact that  $\alpha_k \leq \alpha_0$   
709 for all  $k$  by construction of Algorithm 3.1, and (3.12) that each  $k \geq \bar{k}$  yields

$$\begin{aligned} 710 \quad \tau_{k,\text{trial}} &= \frac{(1 - \sigma_c)(\|c_k\|_2 - \|c_k + J_k v_k\|_2)}{g_k^T s_k + \frac{1}{2\alpha_k} \|s_k\|_2^2 + r(x_k + s_k) - r_k} \\ 711 \quad &\geq \frac{(1 - \sigma_c)\kappa_1(\sigma_{\min}^{\mathcal{L}})^2 \|c_k\|_2 \min\left\{\frac{1}{1+\kappa_J^2}, \kappa_v \alpha_k\right\}}{2(\kappa_{\nabla f} + \kappa_{\partial r})\kappa_v \kappa_J \alpha_k \|c_k\|_2 + 2\kappa_v^2 \kappa_J^2 \kappa_c \alpha_k \|c_k\|_2} \\ 712 \quad &\geq \frac{(1 - \sigma_c)\kappa_1(\sigma_{\min}^{\mathcal{L}})^2 \min\left\{\frac{1}{(1+\kappa_J^2)\alpha_0}, \kappa_v\right\}}{2(\kappa_{\nabla f} + \kappa_{\partial r})\kappa_v \kappa_J + 2\kappa_v^2 \kappa_J^2 \kappa_c}, \end{aligned}$$

713 which proves that (5.29) holds for all  $k \geq \bar{k}$ , as claimed. The merit parameter update  
714 rule (3.10) and (5.29) give (5.30). Finally, (5.31) follows from (5.30) and Lemma 5.6.  $\square$

715 The next result establishes that the norm of the search direction converges to zero  
716 along the sequence of successful iterations.

717 **LEMMA 5.23.** *The search direction sequence  $\{s_k\}_{k \in \mathcal{S}}$  satisfies  $\lim_{k \in \mathcal{S}} \|s_k\|_2 = 0$ .*

718 *Proof.* We first note that the derivation of (5.20) still holds under the assumptions  
719 of this section, and therefore we know that

$$720 \quad (5.32) \quad \bar{\Phi}_{\tau_k}(x_k) - \bar{\Phi}_{\tau_{k+1}}(x_{k+1}) \geq \sum_{k \in \mathcal{S}} \eta_{\Phi} \frac{\tau_k}{8\alpha_k} \|s_k\|_2^2.$$

721 Using nonnegativity of  $\bar{\Phi}_{\tau}$  in (5.7), Lemma 5.7(ii)-(iii), and (5.32), we have that

$$722 \quad \infty > \sum_{k \in \mathcal{S}} (\bar{\Phi}_{\tau_k}(x_k) - \bar{\Phi}_{\tau_{k+1}}(x_{k+1})) \geq \sum_{k \in \mathcal{S}} \eta_{\Phi} \frac{\tau_k}{8\alpha_k} \|s_k\|_2^2.$$

723 Lemma 5.22 gives  $\tau_k \geq \bar{\tau}_{\min} > 0$  for all  $k \in \mathbb{N}$ , where  $\bar{\tau}_{\min}$  is defined in (5.30), so that  
724  $\sum_{k \in \mathcal{S}} \eta_{\Phi} \frac{\bar{\tau}_{\min}}{8\alpha_0} \|s_k\|_2^2 < \infty$ , which implies  $\lim_{k \in \mathcal{S}} \|s_k\|_2 = 0$ , and completes the proof.  $\square$

725 We next prove that the sequence of Lagrange multiplier estimates generated by  
726 subproblem (3.9) during successful iterations are bounded.

727 **LEMMA 5.24.** *There exists  $\kappa_{yz} \in \mathbb{R}_{>0}$  so that  $\max_{k \in \mathcal{S}} \max\{\|y_k\|_{\infty}, \|z_k\|_{\infty}\} \leq \kappa_{yz}$ .*

728 *Proof.* Let  $\bar{k}_1$  serve the role of  $\bar{k}$  in Lemma 5.19 so that the results of Lemma 5.19  
729 hold for each  $k \geq \bar{k}_1$ . Let  $\bar{k}_2$  be sufficiently large so that  $\|s_k\|_2 \leq \frac{1}{3}\delta_{\min}^{\mathcal{L}}$  for all  
730  $\bar{k}_2 \leq k \in \mathcal{S}$ , which is possible because of how  $\delta_{\min}^{\mathcal{L}}$  is defined and Lemma 5.23.

731 For the remainder of the proof, consider an arbitrary  $k$  with  $\max\{\bar{k}_1, \bar{k}_2\} \leq k \in \mathcal{S}$ .  
732 Let  $j \in [n_{\mathcal{L}}]$  be the value guaranteed by Lemma 5.19 to exist so (5.23a)–(5.23c) hold.

Next, consider  $i \in \mathcal{I}(x_j^{\mathcal{L}})$ . It follows from (5.23a), the triangle inequality, the  
definition of  $\bar{k}_2$ , and the definition of  $\delta_{\min}^{\mathcal{L}}$  (see (5.24)) that

$$\|x_k + s_k - x_j^{\mathcal{L}}\|_2 \leq \|x_k - x_j^{\mathcal{L}}\|_2 + \|s_k\|_2 \leq \frac{1}{3}\delta_{\min}(x_j^{\mathcal{L}}) + \frac{1}{3}\delta_{\min}^{\mathcal{L}} \leq \frac{2}{3}\delta_{\min}(x_j^{\mathcal{L}}).$$

This inequality, the definition of  $\delta_{\min}(x_j^{\mathcal{L}})$  (see (5.21)), and  $i \in \mathcal{I}(x_j^{\mathcal{L}})$  imply that

$$[x_k + s_k]_i \geq [x_j^{\mathcal{L}}]_i - \frac{2}{3}\delta_{\min}(x_j^{\mathcal{L}}) \geq \delta_{\min}(x_j^{\mathcal{L}}) - \frac{2}{3}\delta_{\min}(x_j^{\mathcal{L}}) = \frac{1}{3}\delta_{\min}(x_j^{\mathcal{L}}) > 0,$$

733 so that  $i \in \mathcal{I}(x_k + s_k)$ . Thus,  $\mathcal{I}(x_j^{\mathcal{L}}) \subseteq \mathcal{I}(x_k + s_k)$ , or equivalently  $\mathcal{A}(x_k + s_k) \subseteq \mathcal{A}(x_j^{\mathcal{L}})$ .

734 Now, let us introduce the notation  $\mathcal{A}_k^s = \mathcal{A}(x_k + s_k)$ . It follows from  $s_k = v_k + u_k$ ,  
 735 (4.8a),  $[z_k]_i = 0$  for all  $i \notin \mathcal{A}_k^s$  (see (4.8c)), and  $\mathcal{A}_k^s \subseteq \mathcal{A}(x_j^{\mathcal{L}})$  (see above) that

$$736 \quad g_k + \frac{1}{\alpha_k} s_k + g_{r,k} = [J_k^T, I_{\mathcal{A}_k^s}^T] \begin{bmatrix} y_k \\ (z_k)_{\mathcal{A}_k^s} \end{bmatrix} = [J_k^T, I_{\mathcal{A}(x_j^{\mathcal{L}})}^T] \begin{bmatrix} y_k \\ (z_k)_{\mathcal{A}(x_j^{\mathcal{L}})} \end{bmatrix}.$$

Combining this result with (5.23c) and  $\mathcal{A}_k^s \subseteq \mathcal{A}(x_j^{\mathcal{L}})$  it follows that

$$\left\| g_k + \frac{1}{\alpha_k} s_k + g_{r,k} \right\|_2 \geq \sigma_{\min}^{\mathcal{L}} \left\| \begin{bmatrix} y_k \\ (z_k)_{\mathcal{A}(x_j^{\mathcal{L}})} \end{bmatrix} \right\|_2 = \sigma_{\min}^{\mathcal{L}} \left\| \begin{bmatrix} y_k \\ z_k \end{bmatrix} \right\|_2.$$

Combining this inequality with the triangle inequality, (3.12),  $\|s_k\|_2 \leq \frac{1}{3}\delta_{\min}^{\mathcal{L}}$ , and  $\alpha_k \geq \bar{\alpha}_{\min}$  (see (5.31)) it follows that

$$\left\| \begin{bmatrix} y_k \\ z_k \end{bmatrix} \right\|_2 \leq \frac{1}{\sigma_{\min}^{\mathcal{L}}} (\kappa_{\nabla f} + \frac{\delta_{\min}^{\mathcal{L}}}{3\bar{\alpha}_{\min}} + \kappa_{\partial r}).$$

737 Since the right-hand side of this inequality is a constant and independent of  $k$ , we know  
 738 that the sequence of Lagrange multipliers over the successful iterations is bounded.  $\square$

739 **THEOREM 5.25.** *Let Assumption 3.1 and Assumption 5.3 hold. Any limit point  
 740  $x_*$  of the sequence  $\{x_k\}_{k \in \mathcal{S}}$  is a first-order KKT point for problem (1.1).*

741 *Proof.* Let  $x_*$  be a limit point of  $\{x_k\}_{k \in \mathcal{S}}$ , i.e., there exists infinite  $\mathcal{K}_1 \subseteq \mathcal{S}$   
 742 satisfying  $\{x_k\}_{k \in \mathcal{K}_1} \rightarrow x_*$ . From Lemma 5.16 and Lemma 5.20, we have that

$$743 \quad (5.33) \quad 0 = \lim_{k \rightarrow \infty} \|v_k(1)\|_2 \geq \lim_{k \rightarrow \infty} \sigma_{\min}^{\mathcal{L}} \|c_k\|_2 \geq 0,$$

744 which implies that  $0 = \lim_{k \rightarrow \infty} \|c_k\|_2 = \lim_{k \in \mathcal{K}_1} \|c_k\|_2$ . Combining this with continuity  
 745 of  $c$  and  $\{x_k\}_{k \in \mathcal{S}} \rightarrow x_*$  it follows that  $c(x_*) = 0$ .

Next, Lemma 5.24 ensures the existence of a vector pair  $(y_*, z_*) \in \mathbb{R}^m \times \mathbb{R}^n$  and infinite subsequence  $\mathcal{K}_2 \subseteq \mathcal{K}_1$  such that  $\{(y_k, z_k)\}_{k \in \mathcal{K}_2} \rightarrow (y_*, z_*)$ . Also, it follows from Lemma 5.23 and Lemma 4.6 that

$$0 = \lim_{k \in \mathcal{K}_2} \|s_k\|_2 \geq \lim_{k \in \mathcal{K}_2} \|\min\{x_k, -z_k\}\|_2 \geq 0,$$

746 which implies that  $\lim_{k \in \mathcal{K}_2} \|\min\{x_k, -z_k\}\|_2 = 0$ . Combining this with the continuity  
 747 of the  $\min$  operator and  $\{(y_k, z_k)\}_{k \in \mathcal{K}_2} \rightarrow (y_*, z_*)$  it follows that  $\min\{x_*, -z_*\} = 0$ .

It follows from Lemma 5.23 and (5.31) that  $\lim_{k \in \mathcal{K}_2} (1/\alpha_k) \|s_k\|_2 = 0$ . This fact, (4.8a),  $\{(x_k, y_k, z_k)\}_{k \in \mathcal{K}_2} \rightarrow (x_*, y_*, z_*)$ , and continuity of  $g$  and  $J$  give

$$g_{r,*} := -g(x_*) - J(x_*)^T y_* - z_* = \lim_{k \in \mathcal{K}_3} (-g_k - J_k^T y_k - z_k) = \lim_{k \in \mathcal{K}_3} g_{r,k},$$

748 so that  $g(x_*) + g_{r,*} + J(x_*)^T y_* + z_* = 0$ . It follows from this equality,  $c(x_*) = 0$ , and  
 749  $\min\{x_*, -z_*\} = 0$  that  $x_*$  is a first-order KKT point for problem (1.1), as claimed.  $\square$

750 **5.3. Active set Identification.** Our result in this section shows, under suitable  
 751 assumptions, that our method can successfully identify the optimal active set.

752 **THEOREM 5.26.** *Let  $x_*$  be a first-order KKT point for problem (1.1) with La-  
 753 grange multiplier vectors  $y_* \in \mathbb{R}^m$  and  $z_* \in \mathbb{R}_{\leq 0}^n$  for the equality constraints and  
 754 bound constraints, respectively. Suppose that strict complementarity holds, i.e., that  
 755  $\max\{x_*, -z_*\} > 0$ . Let  $\mathcal{S}_1 \subseteq \mathcal{S}$  be such that  $\{x_k\}_{k \in \mathcal{S}_1} \rightarrow x_*$ ,  $\{s_k\}_{k \in \mathcal{S}_1} \rightarrow 0$ , and  
 756  $\{z_k\}_{k \in \mathcal{S}_1} \rightarrow z_*$ . Then,  $\mathcal{A}(x_{k+1}) = \mathcal{A}(x_*)$  for all sufficiently large  $k \in \mathcal{S}_1$ .*

757 *Proof.* We have from the optimality conditions in (4.8) that

758 (5.34)  $\|\min\{x_k + s_k, -z_k\}\|_2 = 0 \text{ for all } k \in \mathbb{N}.$

759 It follows from strict complementarity that  $\epsilon := \min\{[-z_*]_j : j \in \mathcal{A}(x_*)\} > 0$ . Com-  
760 bining this with  $\{z_k\}_{k \in \mathcal{S}_1} \rightarrow z_*$  gives the existence of  $\bar{k} \in \mathbb{N}$  such that  $\|z_k - z_*\|_\infty < \epsilon/2$   
761 for all  $\bar{k} \leq k \leq \mathcal{S}_1$ . Thus, all  $\bar{k} \leq k \in \mathcal{S}_1$  and  $j \in \mathcal{A}(x_*)$  satisfy  $[-z_k]_j > \frac{\epsilon}{2}$ . Combining  
762 this with (5.34) shows that  $[x_{k+1}]_i = [x_k + s_k]_i = 0$  for all  $\bar{k} \leq k \in \mathcal{S}_1$  and  $i \in \mathcal{A}(x_*)$ .  
763 Finally, it follows from  $\{x_k\}_{k \in \mathcal{S}_1} \rightarrow x_*$  and  $\{s_k\}_{k \in \mathcal{S}_1} \rightarrow 0$  that  $[x_{k+1}]_i = [x_k + s_k]_i > 0$   
764 for all  $i \notin \mathcal{A}(x_*)$  and  $k \in \mathcal{S}_1$  sufficiently large, which completes the proof.  $\square$

765 **5.4. Manifold Identification.** In this section, we establish a manifold iden-  
766 tification property for Algorithm 3.1 under certain assumptions. For the definition  
767 of a  $C^2$ -smooth manifold  $\mathcal{M} \subset \mathbb{R}^n$  at a given point in  $\mathbb{R}^n$ , see [37, Definition 2.3].  
768 Our result assumes that the regularizer  $r$  is partly smooth relative to a manifold at a  
769 first-order KKT point; see [37, Definition 3.2].

770 To motivate our assumption that the regularizer is partly smooth, consider  $r(x) =$   
771  $\|x\|_1$  and  $x_* \in \mathbb{R}^n \setminus \{0\}$ . Define the set  $\mathcal{M} = \{x \in \mathbb{R}^n : \text{sgn}(x_i) = \text{sgn}([x_*]_i) \text{ for } i \in$   
772  $\mathcal{I}(x_*)$ , and  $x_i = 0$  for  $i \in \mathcal{A}(x_*)\}$ , which is a  $(|\mathcal{I}(x_*)|)$ -dimensional  $C^2$ -smooth mani-  
773 fold around the point  $x_*$ . Then,  $r$  is partly smooth at  $x_*$  relative to  $\mathcal{M}$ .

774 We are now ready to present our manifold identification property of Algorithm 3.1.  
775 The proof borrows ideas from [35, Lemma 1] and relies on [37, Theorem 4.10].

776 **THEOREM 5.27.** *Let  $x_*$  be a first-order KKT point to problem (1.1) with Lagrange  
777 multiplier vectors  $y_*$  and  $z_*$ , and suppose that  $r$  is convex and partly smooth at  $x_*$   
778 relative to a  $C^2$ -smooth manifold  $\mathcal{M}$ . Assume that the proximal parameter sequence  
779  $\{\alpha_k\}_{k \in \mathbb{N}}$  is bounded away from zero, that there exists a subsequence  $\mathcal{S}_1 \subseteq \mathcal{S}$  such that  
780  $\{(x_k, s_k, y_k, z_k)\}_{k \in \mathcal{S}_1} \rightarrow (x_*, 0, y_*, z_*)$ , and that the non-degeneracy condition*

781 (5.35)  $0 \in \{g(x_*) + J(x_*)^T y_* + z_*\} + \text{relint}(\partial r(x_*))$

782 holds, where *relint* denotes the relative interior of a convex set. Then, it follows that  
783  $x_{k+1} \in \mathcal{M}$  for all sufficiently large  $k \in \mathcal{S}_1$ .

784 *Proof.* Let us define  $\bar{y} = -(g(x_*) + J(x_*)^T y_* + z_*)$ , and note from (5.35) that  
785  $\bar{y} \in \text{relint}(\partial r(x_*))$ . Next, since  $r$  is convex, it is prox-regular [37, Definition 3.6] at  $x_*$   
786 with  $\bar{y}$ . It also follows from  $r$  being convex (thus continuous),  $\{x_k\}_{k \in \mathcal{S}_1} \rightarrow x_*$ , and  
787  $\{s_k\}_{k \in \mathcal{S}_1} \rightarrow 0$  that  $\{x_k + s_k\}_{k \in \mathcal{S}_1} \rightarrow x_*$  and  $\{r(x_k + s_k)\}_{k \in \mathcal{S}_1} \rightarrow r(x_*)$ . Combining  
788 these observations with the assumption in the statement of the theorem that  $r$  is partly  
789 smooth at  $x_*$  relative to a  $C^2$ -smooth manifold  $\mathcal{M}$ , means that every assumption  
790 in [37, Theorem 4.10] holds (with  $r$  and  $x_*$  here playing the role of  $f$  and  $\bar{x}$  in [37,  
791 Theorem 4.10]). To use [37, Theorem 4.10]) to establish our manifold identification  
792 result, it remains to prove that  $\{\text{dist}(\bar{y}, \partial r(x_k + s_k))\}_{k \in \mathcal{S}_1} \rightarrow 0$ , as we now show.

793 It follows from the triangle inequality, (3.12), and (3.13) that

794 (5.36) 
$$\begin{aligned} & \|J(x_k + s_k)^T y_* - J(x_k)^T y_k\|_2 \\ & \leq \|J(x_k + s_k)^T y_* - J(x_k)^T y_* + J(x_k)^T y_* - J(x_k)^T y_k\|_2 \\ & \leq L_J \|s_k\|_2 \|y_*\|_2 + \kappa_J \|y_k - y_*\|_2 \text{ for all } k \in \mathbb{N}. \end{aligned}$$

795 Using (4.8a),  $g_{r,k} \in \partial r(x_k + s_k)$ , (3.12), and (5.36), we have that

$$\begin{aligned}
& \text{dist}(-g(x_k + s_k) - J(x_k + s_k)^T y_* - z_*, \partial r(x_k + s_k)) \\
& \leq \| -g(x_k + s_k) - J(x_k + s_k)^T y_* - z_* - g_{r,k} \|_2 \\
796 & = \left\| g(x_k + s_k) - g(x_k) + (J(x_k + s_k)^T y_* - J(x_k)^T y_k) + (z_* - z_k) - \frac{1}{\alpha_k} s_k \right\|_2 \\
& \leq \| g(x_k + s_k) - g(x_k) \|_2 + \| J(x_k + s_k)^T y_* - J(x_k)^T y_k \|_2 + \| z_* - z_k \|_2 + \frac{1}{\alpha_k} \| s_k \|_2 \\
& \leq L_g \| s_k \|_2 + L_J \| s_k \|_2 \| y_* \|_2 + \kappa_J \| y_k - y_* \|_2 + \| z_k - z_* \|_2 + \frac{1}{\alpha_k} \| s_k \|_2 \quad \text{for all } k \in \mathbb{N}.
\end{aligned}$$

797 This inequality,  $\{(x_k, s_k, y_k, z_k)\}_{k \in \mathcal{S}_1} \rightarrow (x_*, 0, y_*, z_*)$ , and  $\{\alpha_k\}$  bounded from 0 give

$$798 (5.37) \quad \{\text{dist}(-g(x_k + s_k) - J(x_k + s_k)^T y_* - z_*, \partial r(x_k + s_k))\}_{k \in \mathcal{S}_1} \rightarrow 0.$$

799 Next, for all  $k \in \mathbb{N}$ , it follows from [15, Theorem 6.2] that

$$\begin{aligned}
800 & |\text{dist}(\bar{y}, \partial r(x_k + s_k)) - \text{dist}(-g(x_k + s_k) - J(x_k + s_k)^T y_* - z_*, \partial r(x_k + s_k))| \\
801 & \leq \| \bar{y} + g(x_k + s_k) + J(x_k + s_k)^T y_* + z_* \|_2,
\end{aligned}$$

802 which immediately implies that

$$\begin{aligned}
803 & \text{dist}(\bar{y}, \partial r(x_k + s_k)) \leq \text{dist}(-g(x_k + s_k) - J(x_k + s_k)^T y_* - z_*, \partial r(x_k + s_k)) \\
804 & \quad + \| \bar{y} + g(x_k + s_k) + J(x_k + s_k)^T y_* + z_* \|_2.
\end{aligned}$$

805 Combining this inequality with (5.37),  $\{(x_k, s_k, y_k, z_k)\}_{k \in \mathcal{S}_1} \rightarrow (x_*, 0, y_*, z_*)$ , and  
806 continuity of  $g$  and  $J$  shows that  $\{\text{dist}(\bar{y}, \partial r(x_k + s_k))\}_{k \in \mathcal{S}_1} \rightarrow 0$ , which was our goal.  
807 We can now apply [37, Theorem 4.10] to conclude that  $x_k + s_k \in \mathcal{M}$  for all sufficiently  
808 large  $k \in \mathcal{S}_1$ . Since  $x_{k+1} = x_k + s_k$  for all  $k \in \mathcal{S}_1$ , the proof is completed.  $\square$

809 **6. Numerical Results.** We present results from numerical experiments con-  
810 ducted using our Python implementation of Algorithm 3.1. The test problems employ  
811 the  $\ell_1$  regularizer, a widely adopted choice to induce sparse solutions. Our numerical  
812 evaluation has two primary objectives: to demonstrate the numerical performance of  
813 our method using standard optimization metrics, and to assess its capability to cor-  
814 rectly identify the zero-nonzero structure of the solution. Our test problems include  
815 special instances of  $\ell_1$ -regularized optimization problems from the CUTEst [23] test  
816 environment, and instances of sparse canonical correlation analysis.

817 **6.1. Implementation details.** Given  $v_k^c$  in (3.6) as the Cauchy point for sub-  
818 problem (3.1), to find a  $v_k$  satisfying the conditions in (3.4), we first compute

$$819 (6.1) \quad v_k^\infty := \arg \min_{v \in \mathbb{R}^n} m_k(v) \quad \text{s.t. } \|v\|_\infty \leq \kappa_v^\infty \alpha_k \delta_k, \quad x_k + v \in \Omega$$

with  $\kappa_v^\infty \in \mathbb{R}_{>0}$ , which differs from (3.1) only in its use of the infinity-norm. Our motivation for using subproblem (6.1) is that the feasible region only consists of simple bound constraints, which can be handled efficiently by solvers. As long as  $\kappa_v^\infty \leq \frac{1}{\sqrt{n}} \kappa_v$  (which we choose to hold), the solution  $v_k^\infty$  to (6.1) satisfies  $\|v_k^\infty\|_2 \leq \sqrt{n} \|v_k^\infty\|_\infty \leq \sqrt{n} \kappa_v^\infty \alpha_k \delta_k \leq \kappa_v \alpha_k \delta_k$ , meaning that  $v_k^\infty$  satisfies the first two conditions in (3.4). To ensure that the third condition is also satisfied, we set

$$v_k \leftarrow \begin{cases} v_k^c & \text{if } m_k(v_k^c) < m_k(v_k^\infty), \\ v_k^\infty & \text{otherwise.} \end{cases}$$

820 To solve subproblem (6.1), we use the barrier method in Gurobi version 11.0.3 [24].  
 821 Next, to solve subproblem (3.9) (as needed in Line 12 of Algorithm 3.1), we exploit  
 822 the structure of the  $\ell_1$ -norm. By introducing variables  $(p, q) \in \mathbb{R}_{\geq 0}^n \times \mathbb{R}_{\geq 0}^n$  and using  
 823  $e$  to denote a ones vector of appropriate dimension, we solve the equivalent problem

$$824 \quad (6.2) \quad \begin{aligned} & \min_{(u,p,q) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n} g_k^T u + \frac{1}{2\alpha_k} \|u\|_2^2 + \frac{1}{\alpha_k} v_k^T u + \lambda e^T (p + q) \\ & \text{s.t. } J_k u = 0, \quad x_k + v_k + u \in \Omega, \quad p \geq 0, \quad q \geq 0. \end{aligned}$$

825 Problem (6.2) is a convex QP that we solve using the dual active-set QP solver  
 826 in Gurobi. In Algorithm 3.1, the proximal parameter  $\alpha_k$  remains unchanged, i.e.,  
 827  $\alpha_{k+1} \leftarrow \alpha_k$  (Line 19), whenever the sufficient decreasing condition at Line 18 is  
 828 satisfied; in our implementation, we instead update it as  $\alpha_{k+1} \leftarrow \max\{\xi^{-1}\alpha_k, 10\}$ ,  
 829 which allows the proximal parameter to possibly take larger values. We found this  
 830 update strategy to work better in our testing, all of the analysis of Section 5.2.3  
 831 still holds, and the analysis of Section 5.2.2 still holds if this modified update is only  
 832 allowed a finite (possibly large) number of times.

833 The parameters used and initial proximal parameter value are presented in Ta-  
 834 ble 6.1. The starting point  $x_0$  and initial proximal-parameter value  $\alpha_0$  used for the  
 835 test problems will be specified in Section 6.2–6.3.

TABLE 6.1  
*Parameters used by Algorithm 3.1. Recall that  $\kappa_v^\infty$  appears in (6.1).*

$\tau_{-1}$	$\kappa_v$	$\kappa_v^\infty$	$\sigma_c$	$\epsilon_\tau$	$\xi$	$\gamma$	$\eta_\Phi$	$\eta_m$
1	$10^3$	$10^{-2}$	0.1	0.1	0.5	0.5	$10^{-4}$	$10^{-4}$

836 Algorithm 3.1 is terminated when one of the following conditions is satisfied.

- **Approximate KKT point.** Algorithm 3.1 is terminated during the  $k$ th iteration with  $x_k$  considered an approximate KKT point if  $\|c_k\|_2 \leq 10^{-6}$ ,  
 $\|g_k + g_{r,k} + J_k^T y_k + z_k\|_2 \leq 10^{-4}$ , and  $\|\min\{x_k, -z_k\}\|_2 \leq 10^{-4}$ .

- **Time limit.** Algorithm 3.1 is terminated if the running time exceeds 1 hour.

841 As is common in the literature, we scale the problem functions. In particular, the  
 842 objective and its gradient are scaled by the scaling factor

$$843 \quad (6.3) \quad \text{scale\_factor} = \begin{cases} \frac{100}{\|\nabla f(x_0)\|_\infty} & \text{if } \|\nabla f(x_0)\|_\infty > 100, \\ 1 & \text{otherwise.} \end{cases}$$

844 A similar scaling strategy is applied to each constraint  $c_i$  for  $1 \in [m]$ .

845 For comparison, we consider the solver Bazinga,<sup>1</sup> which is a safeguarded aug-  
 846 mented Lagrangian method and, to the best of our knowledge, the only open source  
 847 code that can solve problem (1.1); see [18] for more details. The Bazinga algorithm  
 848 is terminated when one of the following conditions is satisfied.

- **Approximate KKT point.** Bazinga is terminated if a certain primal feasibility and dual stationarity measure are less than  $10^{-6}$ .
- **Not a number.** Bazinga is terminated if a NaN occurs.
- **Time limit.** Algorithm 3.1 is terminated if the running time exceeds 1 hour.

<sup>1</sup>The code package of Bazinga is downloaded from <https://github.com/aldma/Bazinga.jl>

853     **6.2. CUTEst test problems.** We first conduct experiments on a subset of the  
 854     CUTEst test problems. Given the objective function  $f$ , equality constraint  $c_E(x) = 0$ ,  
 855     inequality constraints  $c_l \leq c_I(x) \leq c_u$  for some constant vectors  $c_l$  and  $c_u$ , and bound  
 856     constraints  $b_l \leq x \leq b_u$  for some constant vectors  $b_l$  and  $b_u$  all supplied by CUTEst  
 857     for a given test problem, we solve the  $\ell_1$ -regularized optimization problem

858     (6.4)     
$$\min_{(x,s,a) \in \mathbb{R}^{n+m_I+m}} f(x) + \lambda \|a\|_1 \text{ s.t. } \begin{bmatrix} c_E(x) \\ c_I(x) - s \end{bmatrix} + a = 0, \quad \begin{bmatrix} b_l \\ c_l \end{bmatrix} \leq \begin{bmatrix} x \\ s \end{bmatrix} \leq \begin{bmatrix} b_u \\ c_u \end{bmatrix},$$

859     where  $m_I$  is the number of inequality constraints and  $\lambda \in \mathbb{R}_{>0}$  is a regularization  
 860     parameter. The slack vector  $s$  is introduced to reformulate inequality constraints as  
 861     equality constraints plus bound constraints. The vector  $a$  is introduced in this manner  
 862     so that we can control its sparsity for illustrative purposes in our experiments.

863     The subset of CUTEst problems were chosen based on the following selection  
 864     criteria: (i) the objective function is not constant; (ii) the number of variables and  
 865     constraints satisfy  $1 \leq m \leq n \leq 100$ ; (iii) the total number of inequality constraints  
 866     satisfies  $m_I \geq 1$ . For the choice of  $\lambda$ , we consider the following optimization problem

867     (6.5)     
$$\min_{x \in \mathbb{R}^n, s \in \mathbb{R}^{m_I}} f(x) \quad \text{s.t.} \quad \begin{bmatrix} c_E(x) \\ c_I(x) - s \end{bmatrix} = 0, \quad \begin{bmatrix} b_l \\ c_l \end{bmatrix} \leq \begin{bmatrix} x \\ s \end{bmatrix} \leq \begin{bmatrix} b_u \\ c_u \end{bmatrix},$$

868     and let  $(\bar{x}, \bar{s})$  be a first-order KKT point of this problem with Lagrange multiplier  
 869      $y_{\text{eq}}$  associated with the equality constraints. Then, if  $\lambda \geq \|y_{\text{eq}}\|_\infty$ , the point  $(\bar{x}, \bar{s}, 0)$   
 870     is a first-order KKT point for the optimization problem (6.4). With this observation,  
 871     we set  $\lambda = \|y_{\text{eq}}\|_\infty + 10$  where  $y_{\text{eq}}$  is computed by solving problem (6.5) using  
 872     IPOPT [50]. Problems that are not successfully solved by IPOPT are removed from  
 873     the test problems. The final subset consisted of 81 CUTEst test problems.

874     For our tests, we set  $\alpha_0 = 10$  and  $x_0$  as the initial point supplied by CUTEst.

875     We compare the performance of Algorithm 3.1 and Bazinga using several metrics;  
 876     the results of our tests can be found in Table 6.2. The meaning of the columns found  
 877     in Table 6.2 are described in the following bullet points.

- **Feasible.** The number of test problems for which the corresponding method terminates at a point with constraint violation less than  $10^{-6}$ . For this metric, we see that the two methods behave similarly, with Algorithm 3.1 achieving approximate feasibility on four more test problem.
- **Feasible, Better Objective.** To understand the meaning of this column, let  $f_{\text{Algorithm 3.1}}$  denote the final objective value returned by Algorithm 3.1 and  $f_{\text{Bazinga}}$  denote the final objective value returned by Bazinga. We then define the relative difference in the returned objective function values as

886     (6.6)     
$$f_{\text{diff}} := \frac{f_{\text{Bazinga}} - f_{\text{Algorithm 3.1}}}{\max(1, |\min(f_{\text{Bazinga}}, f_{\text{Algorithm 3.1}})|)}.$$

887     We say that Algorithm 3.1 (resp., Bazinga) has a better relative objective  
 888     value if  $f_{\text{diff}} \geq 10^{-6}$  (resp.,  $f_{\text{diff}} \leq -10^{-6}$ ). Using this terminology, column  
 889     “Feasible, Better Objective” gives the number of test problems for which both  
 890     algorithms terminated at a point with constraint violation less than  $10^{-6}$  and  
 891     the corresponding method has a better relative objective value. For this  
 892     metric, Algorithm 3.1 outperforms Bazinga on 8 additional problems.

- **Performs Better.** The number of test problems for which the corresponding method either (i) meets the constraint violation tolerance and the other

895 method does not, or (ii) both methods reach the constraint violation tolerance  
 896 and the corresponding method has a better relative objective value (see (6.6)).  
 897 For this metric, Algorithm 3.1 outperforms Bazinga by one problem.  
 898 • ***a* is Zero.** The number of test problems for which the corresponding method  
 899 returns  $a = 0$ . Algorithm 3.1 outperforms Bazinga on this metric, with Algo-  
 900 rithm 3.1 (resp., Bazinga) returning  $a = 0$  on 76 (resp., 55) of the problems.  
 901 • ***a* is Small.** The number of test problems for which the corresponding method  
 902 returns  $\|a\|_\infty \leq 10^{-8}$ , thus indicating that  $a$  is small (possibly equal to zero).  
 903 When comparing this column with column “*a* is Zero”, we see that the only  
 904 difference is that Bazinga returns a small (nonzero) value for  $a$  on one addi-  
 905 tional test problem; the results for Algorithm 3.1 are unchanged.  
 906 • **KKT Found.** The number of test problems for which the corresponding  
 907 method terminates with an approximate KKT point. Algorithm 3.1 outper-  
 908 forms Bazinga with Algorithm 3.1 (resp., Bazinga) returning an approximate  
 909 first-order KKT point on 70 (resp., 58) of the problems tested.

TABLE 6.2

*Algorithm 3.1 versus Bazinga on various performance metrics related to solving problem (6.4).*

Method	Feasible	Feasible, Better Objective	Performs Better	<i>a</i> is Zero	<i>a</i> is Small	KKT Found
Algorithm 3.1	71	13	14	76	76	70
Bazinga	67	5	13	55	56	58

910 We conclude this section by comparing the computational times of Algorithm 3.1  
 911 and Bazinga. Figure 6.1 is a Dolan-Moré performance profile [19] for timings, capped  
 912 at  $t = 1000$ . The results show that Algorithm 3.1 (red line) outperforms Bazinga  
 913 (purple line); see [19] for details on interpreting this figure.

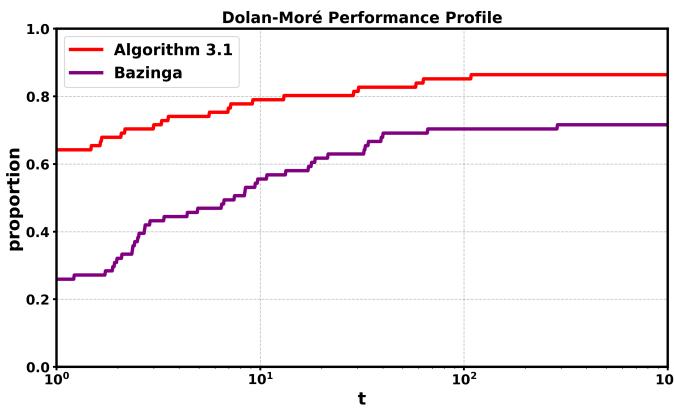


FIG. 6.1. More-Dolen performance profile comparing Algorithm 3.1 and Bazinga in terms of wall-clock time on the subset of CUTEst test problems discussed in Section 6.2.

914        **6.3. Sparse canonical correlation analysis (SCCA).** We now evaluate the  
 915        performance of Algorithm 3.1 on the SCCA problem [52] formulated as

$$916 \quad (6.7) \quad \begin{aligned} \min_{w_x \in \mathbb{R}^{n_x}, w_y \in \mathbb{R}^{n_y}} \quad & -w_x^T \Sigma_{xy} w_y + \lambda(\|w_x\|_1 + \|w_y\|_1) \\ \text{s.t.} \quad & w_x^T \Sigma_{xx} w_x \leq 1, \quad w_y^T \Sigma_{yy} w_y \leq 1, \end{aligned}$$

917        where  $\Sigma_{xx} = XX^T$  and  $\Sigma_{yy} = YY^T$  represent the covariance matrices for data  
 918        matrices  $X \in \mathbb{R}^{n_x \times N}$  and  $Y \in \mathbb{R}^{n_y \times N}$ , respectively, and  $\Sigma_{xy} = XY^T$  represents the  
 919        cross-covariance matrix between  $X$  and  $Y$ . Problem (6.7) aims to identify sparse  
 920        weight vectors  $w_x$  and  $w_y$  that maximize the correlation between the transformed  
 921        views of  $X$  and  $Y$  while the variance constraints prevent trivial solutions where the  
 922        weight vectors are arbitrarily scaled to inflate the correlation.

923        Following the approach of [13], we generate synthetic data matrices  $X$  and  $Y$  as

$$924 \quad X = \left( \begin{bmatrix} e \\ -e \\ 0 \end{bmatrix} + \xi_x \right) u^T \quad \text{and} \quad Y = \left( \begin{bmatrix} 0 \\ e \\ -e \end{bmatrix} + \xi_y \right) u^T,$$

925        where  $e \in \mathbb{R}^{n_x/8}$  represents an all-ones vector,  $\xi_x \in \mathbb{R}^{n_x}$  and  $\xi_y \in \mathbb{R}^{n_y}$  are noise  
 926        vectors with entries sampled from  $\mathcal{N}(0, 0.01)$ , and  $u \in \mathbb{R}^N$  is a random vector with  
 927        entries  $u_i \sim \mathcal{N}(0, 1)$ . This construction creates a known ground truth structure: the  
 928        first  $n_x/4$  rows of  $X$  are correlated with the last  $n_y/4$  rows of  $Y$ . Consequently, the  
 929        ideal sparse solutions for  $w_x$  and  $w_y$  should have non-zero elements confined to the  
 930        first  $n_x/4$  and last  $n_y/4$  indices, respectively.

931        To evaluate the quality of a solution returned by a solver, we compute various  
 932        metrics: the correlation coefficient  $\rho_{xy}$ , sparsity ratio  $sr_x$  for vector  $w_x$ , sparsity ratio  
 933         $sr_y$  for vector  $w_y$ , overall sparsity ratio  $sr$ , variance bound-constraint violations  $voc_x$   
 934        and  $voc_y$ , and sparsity level  $sl$ , which are defined as

$$\begin{aligned} 935 \quad \rho_{xy} &= \frac{w_x^T \Sigma_{xy} w_y}{\sqrt{(w_x^T \Sigma_{xx} w_x)(w_y^T \Sigma_{yy} w_y)}}, \quad sr_x = \frac{n_x - \|w_x\|_0}{n_x}, \\ 936 \quad sr_y &= \frac{n_y - \|w_y\|_0}{n_y}, \quad sr = \frac{(n_x + n_y) - (\|w_x\|_0 + \|w_y\|_0)}{n_x + n_y}, \\ 937 \quad voc_x &= \max(w_x^T \Sigma_{xx} w_x - 1, 0), \quad voc_y = \max(w_y^T \Sigma_{yy} w_y - 1, 0), \text{ and} \\ 938 \quad sl &= \|w_x\|_{[n_x/4+1:n_x]} + \|w_y\|_{[1:3n_y/4-1]}. \end{aligned}$$

939        We consider SCCA test problems of three different sizes with  $n_x = n_y = N \in$   
 940         $\{200, 400, 800\}$  and regularization parameters  $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}\}$ . For each problem  
 941        instance, the starting point  $x_0$  is obtained by solving the generic canonical  
 942        correlation analysis problem (no regularization term) using the `CCA` class from the  
 943        `scikit-learn` package. We set the initial proximal parameter as  $\alpha_0 = 10^{-3}$ . The  
 944        algorithm terminates when one of the conditions detailed in Section 6.1 is satisfied.

945        The results in Table 6.3 demonstrate the effectiveness of Algorithm 3.1 on SCCA  
 946        problems. First, the correlation coefficient achieves the maximum possible value on  
 947        every test case. Second, every solution exhibits the correct sparse structure since  
 948         $sl = 0$ . Third, the algorithm produces solutions with varying sparsity levels that  
 949        are controlled by the regularization parameter  $\lambda$ , with higher sparsity ratios achieved  
 950        by larger  $\lambda$  values. Finally, constraint violations are smaller than  $10^{-9}$ . Table 6.4

TABLE 6.3

Performance metrics for Algorithm 3.1 when solving problem (6.7). Time is measured in seconds.

$n_x = n_y$	$\lambda$	$\rho_{xy}$	$sr_x$	$sr_y$	$sr$	$sl$	$voc_x$	$voc_y$	time
200	$10^{-2}$	1.0000	99.50%	99.50%	99.50%	0	0	0	76.89
	$10^{-3}$	1.0000	99.50%	99.50%	99.50%	0	0	0	87.36
	$10^{-4}$	1.0000	89.50%	90.00%	89.75%	0	0	1.03e-11	117.14
400	$10^{-2}$	1.0000	99.75%	99.75%	99.75%	0	1.40e-9	0	128.40
	$10^{-3}$	1.0000	99.50%	99.00%	99.25%	0	9.83e-11	0	348.44
	$10^{-4}$	1.0000	83.50%	82.75%	83.13%	0	9.46e-11	1.67e-10	226.48
800	$10^{-2}$	1.0000	99.88%	99.88%	99.88%	0	5.86e-9	3.34e-9	279.18
	$10^{-3}$	1.0000	99.63%	99.88%	99.75%	0	6.33e-10	1.81e-9	899.06
	$10^{-4}$	1.0000	96.63%	95.63%	96.13%	0	0	1.47e-10	463.84

TABLE 6.4

Performance metrics for Bazinga when solving problem (6.7). Time is measured in seconds.

$n_x = n_y$	$\lambda$	$\rho_{xy}$	$sr_x$	$sr_y$	$sr$	$sl$	$voc_x$	$voc_y$	time
200	$10^{-2}$	1.0000	99.50%	99.50%	99.50%	0	4.02e-9	3.34e-8	86.10
	$10^{-3}$	1.0000	99.50%	99.50%	99.50%	0	1.96e-8	0	251.97
	$10^{-4}$	1.0000	92.00%	87.50%	89.75%	0	0	0	164.08
400	$10^{-2}$	1.0000	99.75%	99.75%	99.75%	0	6.62e-9	1.32e-8	556.60
	$10^{-3}$	1.0000	97.50%	97.75%	97.63%	0	0	0	744.31
	$10^{-4}$	1.0000	77.75%	85.00%	81.38%	0	0	0	713.13
800	$10^{-2}$	1.0000	98.75%	98.38%	98.56%	0	0	2.35e-9	2958.89
	$10^{-3}$	1.0000	88.63%	97.25%	92.94%	0	0	2.00e-8	2789.95
	$10^{-4}$	1.0000	81.38%	78.75%	80.06%	0	6.55e-8	0	2612.26

951 reports the performance of Bazinga on the same problems. Notably, Algorithm 3.1  
952 attains sparsity ratios that are at least as high as those of Bazinga (sometimes strictly  
953 higher), while requiring less computational time.

954 **7. Conclusion.** We presented the first proximal-gradient-type method for reg-  
955 ularized optimization problems with general nonlinear inequality constraints. Simi-  
956 lar to the traditional proximal-gradient method, we proved that our approach has a  
957 convergence result (under an LICQ assumption), a worst-case iteration complexity  
958 result (under a stronger assumption), as well as a manifold identification property  
959 and active-set identification property (under standard assumptions).

960

## REFERENCES

961 [1] Yanqin Bai, Renli Liang, and Zhouwang Yang. Splitting augmented Lagrangian method for  
962 optimization problems with a cardinality constraint and semicontinuous variables. *Opti-*  
963 *mization Methods and Software*, 31(5):1089–1109, 2016.

964 [2] Gilles Bareilles, Franck Iutzeler, and Jérôme Malick. Newton acceleration on manifolds identi-  
965 fied by proximal gradient methods. *Mathematical Programming*, 200(1):37–70, 2023.

966 [3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse  
967 problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

968 [4] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

969 [5] Carla Bertocchi, Emilie Chouzenoux, Marie-Caroline Corbineau, Jean-Christophe Pesquet, and  
970 Marco Prato. Deep unfolding of a proximal interior point method for image restoration.  
971 *Inverse Problems*, 36(3):034005, 2020.

972 [6] Dimitri Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific, 2009.

973 [7] Digvijay Boob, Qi Deng, and Guanghui Lan. Level constrained first order methods for function  
974 constrained optimization. *Mathematical Programming*, 209(1):1–61, 2025.

975 [8] Lahcen El Bourkhissi, Ion Necoara, Panagiotis Patrinos, and Quoc Tran-Dinh. Complexity of  
 976 linearized perturbed augmented Lagrangian methods for nonsmooth nonconvex optimization  
 977 with nonlinear equality constraints. *arXiv preprint arXiv:2503.01056*, 2025.

978 [9] Paul H. Calamai and Jorge J. Moré. Projected gradient methods for linearly constrained  
 979 problems. *Mathematical programming*, 39(1):93–116, 1987.

980 [10] Tianyi Chen, Frank E. Curtis, and Daniel P. Robinson. A reduced-space algorithm for mini-  
 981 mizing  $\ell_1$ -regularized convex functions. *SIAM J. Optim.*, 27(3):1583–1610, 2017.

982 [11] Tianyi Chen, Frank E. Curtis, and Daniel P. Robinson. FaRSA for  $\ell_1$ -regularized convex opti-  
 983 mization: local convergence and numerical experience. *Optim. Methods. Softw.*, 33(2):396–  
 984 415, 2018.

985 [12] Emilie Chouzenoux, Marie-Caroline Corbineau, and Jean-Christophe Pesquet. A proximal  
 986 interior point algorithm with applications to image processing. *Journal of Mathematical  
 987 Imaging and Vision*, 62(6):919–940, 2020.

988 [13] Delin Chu, Li-Zhi Liao, Michael K Ng, and Xiaowei Zhang. Sparse canonical correlation analy-  
 989 sis: New formulation and algorithm. *IEEE transactions on pattern analysis and machine  
 990 intelligence*, 35(12):3050–3065, 2013.

991 [14] XT Cui, XJ Zheng, SS Zhu, and XL Sun. Convex relaxations and MIQCQP reformulations  
 992 for a class of cardinality-constrained portfolio selection problems. *Journal of Global Opti-  
 993 mization*, 56(4):1409–1423, 2013.

994 [15] Frank E. Curtis and Daniel P. Robinson. *Practical Nonconvex Nonsmooth Optimization*. MOS-  
 995 SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadel-  
 996 phia, PA, USA, 2025.

997 [16] Yutong Dai, Xiaoyi Qu, and Daniel P. Robinson. A proximal-gradient method for equality  
 998 constrained optimization. *SIAM Journal on Optimization*, 35(4):2654–2683, 2025.

999 [17] Alberto De Marchi. An interior proximal gradient method for nonconvex optimization. *Open  
 1000 Journal of Mathematical Optimization*, 5:1–22, 2024.

1001 [18] Alberto De Marchi, Xiaoxi Jia, Christian Kanzow, and Patrick Mehlitz. Constrained com-  
 1002 posite optimization and augmented Lagrangian methods. *Mathematical Programming*,  
 1003 201(1):863–896, 2023.

1004 [19] Elizabeth D. Dolan and Jorge J. Moré. Benchmarking optimization software with performance  
 1005 profiles. *Mathematical programming*, 91(2):201–213, 2002.

1006 [20] Florian Dörfler, Mihailo R Jovanović, Michael Chertkov, and Francesco Bullo. Sparsity-  
 1007 promoting optimal wide-area control of power networks. *IEEE Transactions on Power  
 1008 Systems*, 29(5):2281–2291, 2014.

1009 [21] Francisco Facchinei, Andreas Fischer, and Christian Kanzow. On the accurate identification of  
 1010 active constraints. *SIAM Journal on Optimization*, 9(1):14–32, 1998.

1011 [22] Makan Fardad, Fu Lin, and Mihailo R. Jovanović. Sparsity-promoting optimal control for a  
 1012 class of distributed systems. In *Proceedings of the 2011 American Control Conference*,  
 1013 pages 2050–2055. IEEE, 2011.

1014 [23] Nicholas I.M. Gould, Dominique Orban, and Philippe L. Toint. CUTEst: a constrained and  
 1015 unconstrained testing environment with safe threads for mathematical optimization. *Com-  
 1016 putational optimization and applications*, 60:545–557, 2015.

1017 [24] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023.

1018 [25] Davood Hajinezhad and Mingyi Hong. Perturbed proximal primal–dual algorithm for noncon-  
 1019 vex nonsmooth optimization. *Mathematical Programming*, 176(1):207–245, 2019.

1020 [26] Nadav Hallak and Marc Teboulle. An adaptive Lagrangian-based scheme for nonconvex com-  
 1021 posite optimization. *Mathematics of Operations Research*, 48(4):2337–2352, 2023.

1022 [27] Syed Ali Hamza and Moeness G Amin. Hybrid sparse array beamforming design for general  
 1023 rank signal models. *IEEE Transactions on Signal Processing*, 67(24):6215–6226, 2019.

1024 [28] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections  
 1025 for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

1026 [29] Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in  
 1027 deep learning: Pruning and growth for efficient inference and training in neural networks.  
*Journal of Machine Learning Research*, 22(241):1–124, 2021.

1028 [30] Huiping Huang, Hing Cheung So, and Abdelhak M Zoubir. Sparse array beamformer design  
 1029 via ADMM. *IEEE Transactions on Signal Processing*, 71:3357–3372, 2023.

1030 [31] Bo Jiang, Tianyi Lin, Shiqian Ma, and Shuzhong Zhang. Structured nonconvex and nonsmooth  
 1031 optimization: algorithms and iteration complexity analysis. *Computational Optimization  
 1032 and Applications*, 72(1):115–157, 2019.

1033 [32] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient  
 1034 methods under the Polyak–Lojasiewicz condition. In *Joint European Conference on Ma-  
 1035 chine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

1037 [33] Weiwei Kong, Jefferson G. Melo, and Renato D.C. Monteiro. Complexity of a quadratic penalty  
 1038       accelerated inexact proximal point method for solving linearly constrained nonconvex com-  
 1039       posite programs. *SIAM Journal on Optimization*, 29(4):2566–2593, 2019.

1040 [34] Geoffroy Leconte and Dominique Orban. An interior-point trust-region method for nonsmooth  
 1041       regularized bound-constrained optimization. *arXiv preprint arXiv:2402.18423*, 2024.

1042 [35] Ching-pei Lee. Accelerating inexact successive quadratic approximation for regularized op-  
 1043       timization through manifold identification. *Mathematical Programming*, 201(1):599–633,  
 1044       2023.

1045 [36] Ching-pei Lee and Stephen J. Wright. Inexact successive quadratic approximation for regular-  
 1046       ized optimization. *Comput. Optim. Appl.*, 72:641–674, 2019.

1047 [37] Adrian S. Lewis and Shanshan Zhang. Partial smoothness, tilt stability, and generalized Hes-  
 1048       sians. *SIAM Journal on Optimization*, 23(1):74–94, 2013.

1049 [38] Zichong Li, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Yangyang Xu. Rate-improved inexact  
 1050       augmented Lagrangian method for constrained nonconvex optimization. In *International  
 1051       Conference on Artificial Intelligence and Statistics*, pages 2170–2178. PMLR, 2021.

1052 [39] Shuai Liu, Claudia Sagastizabal, and Mikhail V. Solodov. Proximal gradient-method with su-  
 1053       perlinear convergence for nonsmooth convex optimization. *SIAM Journal on Optimization*,  
 1054       35(3):1601–1629, 2025.

1055 [40] G.P. McCormick and R.A. Tapia. The gradient projection method under mild differentiability  
 1056       conditions. *SIAM Journal on Control*, 10(1):93–98, 1972.

1057 [41] Jorge J Moré. Trust regions and projected gradients. In *System Modelling and Optimization:  
 1058       Proceedings of the 13th IFIP Conference Tokyo, Japan, August 31–September 4, 1987*,  
 1059       pages 1–13. Springer, 2006.

1060 [42] Julie Nutini, Mark Schmidt, and Warren Hare. “Active-set complexity” of proximal gradient:  
 1061       How long does it take to find the sparsity pattern? *Optimization Letters*, 13:645–655,  
 1062       2019.

1063 [43] Christina Oberlin and Stephen J. Wright. Active set identification in nonlinear programming.  
 1064       *SIAM Journal on Optimization*, 17(2):577–605, 2006.

1065 [44] Daniel P. Robinson. *Primal-Dual Methods for Nonlinear Optimization*. PhD thesis, Department  
 1066       of Mathematics, University of California San Diego, La Jolla, CA, 2007.

1067 [45] R. Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science  
 1068       & Business Media, 2009.

1069 [46] Mehmet Fatih Sahin, Ahmet Alacaoglu, Fabian Latorre, Volkan Cevher, et al. An inexact  
 1070       augmented Lagrangian framework for nonconvex optimization with nonlinear constraints.  
 1071       *Advances in Neural Information Processing Systems*, 32, 2019.

1072 [47] Yifan Sun, Halyun Jeong, Julie Nutini, and Mark Schmidt. Are we there yet? manifold  
 1073       identification of gradient-related proximal methods. In *The 22nd International Conference  
 1074       on Artificial Intelligence and Statistics*, pages 1110–1119. PMLR, 2019.

1075 [48] Ph. L. Toint. Global convergence of a of trust-region methods for nonconvex minimization in  
 1076       hilbert space. *IMA Journal of Numerical Analysis*, 8(2):231–252, 1988.

1077 [49] Steve Tonneau, Daeun Song, Pierre Fernbach, Nicolas Mansard, Michel Taïx, and Andrea  
 1078       Del Prete. SL1M: Sparse L1-norm minimization for contact planning on uneven terrain. In  
 1079       *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6604–  
 1080       6610. IEEE, 2020.

1081 [50] Andreas Wächter and Lorenz T. Biegler. On the implementation of an interior-point filter  
 1082       line-search algorithm for large-scale nonlinear programming. *Mathematical programming*,  
 1083       106(1):25–57, 2006.

1084 [51] Pinzheng Wei and Weihong Yang. An SQP-type proximal gradient method for composite  
 1085       optimization problems with equality constraints. *Journal of Computational Mathematics*,  
 1086       2024.

1087 [52] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposi-  
 1088       tion, with applications to sparse principal components and canonical correlation analysis.  
 1089       *Biostatistics*, 10(3):515–534, 2009.

1090 [53] Xinghao Yang, Weifeng Liu, Wei Liu, and Dacheng Tao. A survey on canonical correlation  
 1091       analysis. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2349–2368, 2019.

1092 [54] Yuqian Zhang, Yenson Lau, Han-wen Kuo, Sky Cheung, Abhay Pasupathy, and John Wright.  
 1093       On the global geometry of sphere-constrained sparse blind deconvolution. In *Proceedings  
 1094       of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4902,  
 1095       2017.

1096 [55] Hui Zou and Lingzhou Xue. A selective overview of sparse principal component analysis.  
 1097       *Proceedings of the IEEE*, 106(8):1311–1320, 2018.