Journal of Optimization Theory and Applications
https://doi.org/10.1007/s10957-024-02421-6

Computing Second-Order Points Under Equality Constraints: Revisiting Fletcher's Augmented Lagrangian

Florentin Goyens[1] · Armin Eftekhari[2] · Nicolas Boumal[3]

Received: 2 September 2023 / Accepted: 8 March 2024

Abstract

We address the problem of minimizing a smooth function under smooth equality constraints. Under regularity assumptions on these constraints, we propose a notion of approximate first- and second-order critical point which relies on the geometric formalism of Riemannian optimization. Using a smooth exact penalty function known as Fletcher's augmented Lagrangian, we propose an algorithm to minimize the penalized cost function which reaches ε-approximate second-order critical points of the original optimization problem in at most $\mathcal{O}(\varepsilon^{-3})$ iterations. This improves on current best theoretical bounds. Along the way, we show new properties of Fletcher's augmented Lagrangian, which may be of independent interest.

Keywords Nonconvex optimization · Constrained optimization · Augmented Lagrangian · Complexity · Riemannian optimization

1 Introduction

Working over a Euclidean space $\mathcal{E}$ with inner product $\langle\cdot,\cdot\rangle$ and associated norm $\|\cdot\|$, we consider the constrained optimization problem

$$\min_x f(x) \text{ subject to } h(x) = 0, \tag{P}$$

Communicated by Aram Arutyunov.

✉ Florentin Goyens
goyensflorentin@gmail.com

[1] LAMSADE, Université Paris Dauphine-PSL, Paris, France

[2] Umeå, Sweden

[3] Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Published online: 11 April 2024

where $f : \mathcal{E} \to \mathbb{R}$ and $h : \mathcal{E} \to \mathbb{R}^m$ are smooth (C^∞). The feasible set is denoted by

$$\mathcal{M} = \{x \in \mathcal{E} : h(x) = 0\}. \tag{1.1}$$

Our aim is to propose an infeasible algorithm for problem (P) that has good global complexity guarantees—an active topic of research. The complexity is expressed in terms of worst-case number of iterations needed to find an ε-approximate (second-order) critical point. Thus, we need a precise notion of approximate criticality. For constrained problems such as (P), especially when it comes to second-order criticality, there does not seem to be a consensus on what that should be.[1]

Here, under a certain LICQ-type assumption (see A1 below), we propose a natural notion of ε-approximate second-order optimality conditions in Sect. 1.2. Our definition has a geometric interpretation, as it is an extension of the Riemannian optimality conditions to points that are approximately feasible. This allows us to use the formalism of Riemannian optimization for the complexity analysis of an infeasible method. This perspective, combined with a modern take on some of Fletcher's ideas from the 1970 s, leads to improved complexity bounds.

Concretely, we propose an algorithm which computes such ε-approximate second-order critical points with state-of-the-art worst-case iteration complexity with respect to ε (Sect. 3). The algorithm relies on an augmented Lagrangian formalism introduced by Fletcher [20] which provides a *smooth yet exact* penalty function for constrained optimization given by

$$g(x) := f(x) - \langle h(x), \lambda(x)\rangle + \beta \left\|h(x)\right\|^2, \tag{1.2}$$

for some parameter $\beta \geq 0$ and multipliers $\lambda(x)$ defined below in (1.7). This penalty function has a reputation for being impractical. We study it in its original form as it allows us to secure desirable theoretical guarantees. Moreover, we note that other authors [18, 19, 21] have successfully used approximations of g or its derivatives to build practical schemes, and it is possible that the theoretical guarantees could extend to those as well.

Our theoretical algorithm is a simple method that combines gradient and eigensteps applied to Fletcher's augmented Lagrangian. The existing literature on Fletcher's augmented Lagrangian focuses on asymptotic convergence towards minimizers. We complement this with an analysis of the global complexity of computing approximate critical points of (P) (non-asymptotic). We do so in two phases.

First, we show that approximate critical points of Fletcher's penalty are approximate critical points of (P). This is an extension of known results which relate exact minimizers of g to exact minimizers of (P). Second, we show that our algorithm computes approximate minimizers of g in finite time, and we give a worst-case bound on the number of iterations for their computation. This leads to a complexity rate with respect to ε which improves on the state of the art for computing second-order critical points under equality constraints (even after taking into account the differences in

[1] We review various proposals that have been made, with their pros and cons, in Appendix A of the ArXiv version of this paper [23].

notions of approximate criticality). One downside of the algorithm is that it requires properly setting a penalty parameter β: we discuss how to circumvent this issue in Sect. 4, at the cost of log-factors.

There is a wealth of related literature. We summarize the complexity results in Table 1, and provide further details regarding complexity and notions of approximate criticality in Appendix A of the arXiv version of this paper. For our contributions and outline of the paper, see Sect. 1.4. We preface this with our assumptions on Problem (P) in Sect. 1.1 and our geometric definition of approximate critical points in Sect. 1.2.

1.1 Assumptions

We introduce three central assumptions about the set $\mathcal{M}$ (1.1). The following set is open:

$$\mathcal{D} = \{x \in \mathcal{E} \colon \operatorname{rank}(\mathrm{D}h(x)) = m\}. \tag{1.3}$$

It is known that if $\mathcal{M}$ is included in $\mathcal{D}$ then $\mathcal{M}$ is a (smooth) embedded submanifold of $\mathcal{E}$ [2]. We further assume that there is a region around $\mathcal{M}$ where the differential of the constraints is nonsingular: this is the classical linear independence constraint qualification (LICQ). Below we use $\|\cdot\|$ to denote the 2-norm on $\mathbb{R}^m$.

A 1 There exist positive constants $R, \underline{\sigma}$ such that for all x in the set

$$\mathcal{C} = \{x \in \mathcal{E} : \|h(x)\| \leq R\} \tag{1.4}$$

we have $\sigma_{\min}(\mathrm{D}h(x)) = \sigma_m(\mathrm{D}h(x)) \geq \underline{\sigma} > 0$ where $\sigma_k(A)$ and $\sigma_{\min}(A)$ denote the kth and the smallest singular value of a linear map A, respectively. In particular, $\mathcal{M} \subset \mathcal{C} \subset \mathcal{D}$.

A 2 The sets $\mathcal{M} = \{x \in \mathcal{E} : h(x) = 0\}$ and $\mathcal{C} = \{x \in \mathcal{E} : \|h(x)\| \leq R\}$ are compact.

Rather than assuming that $\mathcal{M}$ and $\mathcal{C}$ are compact, our results could be extended to assume instead that the sublevel set $\{x \in \mathcal{E} \colon g(x) \leq g(x_0)\}$ is bounded. We choose to proceed with A2 to avoid assumptions which would mix h and f (as is the case for g). In doing so, we favor assumptions that can be checked once for a given constraint h, leading to results that apply for broad choices of cost function f.

A 3 There exists a constant $C_h > 0$ such that, for all $x \in \mathcal{C}$ and $v \in \mathcal{E}$,

$$h(x + v) = h(x) + \mathrm{D}h(x)[v] + E(x, v)$$

with $\|E(x, v)\| \leq C_h \|v\|^2$.

Given the nonconvex nature of (P), it is necessary to make some assumption in order to guarantee convergence to a feasible point. The set $\mathcal{C}$ is the region where our assumptions apply. Accordingly, we require initialization in $\mathcal{C}$. In some cases, such initializations are easy to produce (see Stiefel example below); in other cases, one may resort to a two-phase algorithm whose first phase attempts to compute an approximately feasible point [13].

A 4 The iterate x_0 belongs to $\mathcal{C}$.

Note that affine constraints do not satisfy A2, but these constraints are usually not problematic as there are various effective approaches to handle them, including feasible methods. The following example shows how to compute the constants R and $\underline{\sigma}$, which define the region of interest $\mathcal{C}$, for the Stiefel manifold.

Example (*The Stiefel manifold*) Let $\mathcal{E} = \mathbb{R}^{n\times p}$ for $1 \le p \le n$. The Stiefel manifold is defined as

$$\mathrm{St}(n, p) = \{X \in \mathbb{R}^{n\times p} : X^\top X = \mathrm{I}_p\}. \tag{1.5}$$

The manifold corresponds to the defining function $h\colon \mathbb{R}^{n\times p} \to \mathrm{Sym}(p)\colon X \mapsto h(X) = X^\top X - \mathrm{I}_p$, where $\mathrm{Sym}(p)$ is the set of symmetric matrices of size p. For any $R < 1$, it is possible to verify that all $X \in \mathbb{R}^{n\times p}$ such that $\|h(X)\| \le R$ satisfy $\sigma_{\min}(\mathrm{D}h(X)) \ge 2\sigma_{\min}(X) \ge 2\sqrt{1-R}$. Therefore, A1 is satisfied for any $R < 1$ and $\underline{\sigma} \le 2\sqrt{1-R}$. For any $R > 0$, the set $\mathrm{St}(n, p)$ satisfies the compactness assumption A2. Assumption A3 holds with $C_h = 1$. Additionally, A4 is easily satisfied by taking a matrix with p orthonormal columns in $\mathbb{R}^n$ as initial iterate.

Given several sets that satisfy the assumptions above, their Cartesian product also does.

Proposition 1.1 *For $i = 1, 2, \ldots, k$, consider k functions $h_i\colon \mathcal{E}_i \to \mathbb{R}^{m_i}$ that satisfy assumptions A1, A2 and A3 with constants R_i, $\underline{\sigma}_i$ and C_{h_i}. Then, the function $h\colon \mathcal{E} \to \mathbb{R}^m$ with $\mathcal{E} = \mathcal{E}_1 \times \cdots \times \mathcal{E}_k$ and $m = m_1 + \cdots + m_k$ defined by $h(x_1, \ldots, x_k) = (h_1(x_1), \ldots, h_k(x_k))^\top$ satisfies A1, A2 and A3 with constants $R = \min(R_1, \ldots, R_k)$, $\underline{\sigma} = \min(\underline{\sigma}_1, \ldots, \underline{\sigma}_k)$ and $C_h = \max(C_{h_1}, \ldots, C_{h_k})$.*

Proof Let $\mathcal{C} = \{x \in \mathcal{E}\colon \|h(x)\| \le R\}$ and $x = (x_1, \ldots, x_k) \in \mathcal{C}$. Since $\|h(x)\| \le \min(R_1, \ldots, R_k)$, we have $\|h_i(x_i)\| \le R_i$ and $\sigma_{\min}(\mathrm{D}h_i(x_i)) \ge \underline{\sigma}_i$ for all $i = 1, \ldots, k$. This implies $\sigma_{\min}(\mathrm{D}h(x)) \ge \min(\underline{\sigma}_1, \ldots, \underline{\sigma}_k)$. The Cartesian product of compact sets is compact. Let $v = (v_1, \ldots, v_k) \in \mathcal{E}$. Since $v_i \in \mathcal{E}_i$ and $x_i \in \mathcal{C}_i$, we have $h_i(x_i + v_i) = h_i(x_i) + \mathrm{D}h_i(x_i)[v_i] + E_i(x_i, v_i)$ with $\|E_i(x_i, v_i)\| \le C_{h_i}\|v_i\|^2$ for all $i = 1, \ldots, k$. Define $E(x, v) = (E_1(x_1, v_1), \ldots, E_k(x_k, v_k))^\top$, then $h(x + v) = h(x) + \mathrm{D}h(x)[v] + E(x, v)$ with $\|E(x, v)\| \le \sum_{i=1}^k \|E_i(x_i, v_i)\| \le \sum_{i=1}^k C_{h_i}\|v_i\|^2 \le \left(\max_i C_{h_i}\right)\|v\|^2$. □

As the Stiefel manifold includes spheres ($p = 1$) and orthogonal matrices ($p = n$) as special cases, Proposition 1.1 establishes that products of spheres and orthogonal/rotation groups satisfy the assumptions above. This covers a wide range of applications, of which we list a few. A product of spheres appears in the Burer–Monteiro factorization of semidefinite programs with diagonal constraints [11]. It also appears in independent component analysis (ICA) and orthogonal tensor decomposition [22], and models the rank reduction of correlation matrices [25]. Products of orthogonal matrices appear in applications of orthogonal group synchronisation [29]. The simultaneous localization and mapping problem in robotics involves optimization over a product of Stiefel manifolds [34].

1.2 Optimality Conditions on Layered Manifolds

Assumption A1 allows us to characterize any point in $\mathcal{C}$ as belonging to some Riemannian submanifold of $\mathcal{E}$. This manifold is defined by a level set of the function h, while the feasible set $\mathcal{M}$ is the zero-set of h. This observation partitions the region of interest $\mathcal{C}$ into Riemannian submanifolds which we call *layered manifolds*. These layered manifolds help to formulate meaningful criticality conditions for points which are nearly but not exactly feasible.

Proposition 1.2 (Layered manifolds) *Under* A1, *for any $x \in \mathcal{C}$, the set $\mathcal{M}_x = \{y \in \mathcal{E} : h(y) = h(x)\}$ is a submanifold of $\mathcal{E}$ contained in $\mathcal{C}$. The tangent space and the normal space of $\mathcal{M}_x$ at $y \in \mathcal{M}_x$ are given respectively by:*

$$\mathrm{T}_y\mathcal{M}_x = \ker \mathrm{D}h(y) \quad \textit{and} \quad \mathrm{N}_y\mathcal{M}_x = \operatorname{span}\big(\mathrm{D}h(y)^*\big), \tag{1.6}$$

where a star indicates an adjoint.

Proof Using Proposition 3.3.3 from [2], the set $\mathcal{M}_x$ is a submanifold of $\mathcal{E}$ if $\operatorname{rank}(\mathrm{D}h(y)) = m$ for all $y \in \mathcal{M}_x$, which holds for all $x \in \mathcal{C}$ under A1. □

The embedded submanifold $\mathcal{M}_x$ for some $x \in \mathcal{C}$ is turned into a Riemannian submanifold using the Euclidean inner product of $\mathcal{E}$ restricted to the tangent spaces of $\mathcal{M}_x$. We proceed to compute the Riemannian gradient and Riemannian Hessian of f on the layer $\mathcal{M}_x$. To this end, we define the function $\lambda \colon \mathcal{E} \to \mathbb{R}^m$ as follows:

$$\lambda(x) = (\mathrm{D}h(x)^*)^\dagger[\nabla f(x)], \tag{1.7}$$

where a dagger indicates a Moore–Penrose pseudo-inverse. *This is the same function* $\lambda(\cdot)$ *used in Fletcher's augmented Lagrangian* (Eq. 1.2). This function is particularly relevant at points x in $\mathcal{C}$ because, if $\operatorname{rank} \mathrm{D}h(x) = m$, then the orthogonal projector from $\mathcal{E}$ to the tangent space $\mathrm{T}_x\mathcal{M}_x = \ker \mathrm{D}h(x)$ is given in explicit form by

$$\mathrm{Proj}_x(v) = v - \mathrm{D}h(x)^*[z] \quad \text{with} \quad z = (\mathrm{D}h(x)^*)^\dagger[v].$$

Therefore, the Riemannian gradient of f on $\mathcal{M}_x$ is given by

$$\mathrm{grad}_{\mathcal{M}_x} f(x) = \mathrm{Proj}_x(\nabla f(x)) = \nabla f(x) - \mathrm{D}h(x)^*[\lambda(x)], \tag{1.8}$$

the orthogonal projection of the Euclidean gradient of f to the tangent space $\mathrm{T}_x\mathcal{M}_x$. Likewise, the Riemannian Hessian of f on $\mathcal{M}_x$ is given by

$$\mathrm{Hess}_{\mathcal{M}_x} f(x) = \mathrm{Proj}_x \circ \left(\nabla^2 f(x) - \sum_{i=1}^{m} \lambda_i(x)\nabla^2 h_i(x)\right) \circ \mathrm{Proj}_x, \tag{1.9}$$

a self-adjoint linear operator on $\mathrm{T}_x\mathcal{M}_x$ [10, Section 7.7].

We now go over exact and approximate criticality conditions for problem (P). First-order critical points of (P) are defined by

$$h(x) = 0 \quad \text{and} \quad \mathrm{grad}_{\mathcal{M}} f(x) = 0, \tag{1.10}$$

whereas second-order critical points satisfy

$$h(x) = 0, \quad \mathrm{grad}_{\mathcal{M}} f(x) \quad = 0, \quad \text{and} \quad \mathrm{Hess}_{\mathcal{M}} f(x) \succeq 0. \tag{1.11}$$

At points $x \in \mathcal{D}$ (1.3), constraint qualifications hold, providing:

Proposition 1.3 *Any local minimizer of (P) is a second-order critical point which satisfies (1.11).*

Using this Riemannian viewpoint, we propose a new definition of approximate criticality for smooth equality constraints. We compare this new notion to existing ones in [23, Appendix A].

Definition 1.1 The point $x \in \mathcal{D}$ is an $(\varepsilon_0, \varepsilon_1)$-approximate first-order critical point of (P) if

$$\|h(x)\| \leq \varepsilon_0 \quad \text{and} \quad \left\|\mathrm{grad}_{\mathcal{M}_x} f(x)\right\| \leq \varepsilon_1. \tag{ε-FOCP}$$

Definition 1.2 The point $x \in \mathcal{D}$ is an $(\varepsilon_0, \varepsilon_1, \varepsilon_2)$-approximate second-order critical point of (P) if

$$\|h(x)\| \leq \varepsilon_0, \quad \left\|\mathrm{grad}_{\mathcal{M}_x} f(x)\right\| \leq \varepsilon_1 \quad \text{and} \quad \mathrm{Hess}_{\mathcal{M}_x} f(x) \succeq -\varepsilon_2 \mathrm{Id}. \tag{ε-SOCP}$$

The notions of (ε-FOCP) and (ε-SOCP) have a natural geometric interpretation. For a point $x \in \mathcal{C}$ which is nearly feasible, the criticality is assessed with respect to the manifold layer to which x belongs. In essence, x satisfies the usual approximate criticality conditions for a Riemannian optimization problem, i.e., small Riemannian gradient and almost positive semi-definite Riemannian Hessian. However, these conditions are satisfied on the tangent space of a layer manifold $\mathcal{M}_x$ rather than on the target manifold $\mathcal{M}$.

1.3 Related Work

The study of complexity in optimization gives guarantees on the worst-case number of iterations an algorithm requires to achieve a predetermined termination criterion. In the unconstrained case, where $\mathcal{M} = \mathcal{E}$, Nesterov [31] shows that for Lipschitz differentiable f, gradient descent with an appropriate step size requires at most $\mathcal{O}(\varepsilon^{-2})$ iterations to find a point which satisfies $\|\nabla f(x)\| \leq \varepsilon$. Cartis et al. [12] further show that a point which satisfies both $\|\nabla f(x)\| \leq \varepsilon$ and $\lambda_{\min}(\nabla^2 f(x)) \geq -\varepsilon$ can be found in $\mathcal{O}(\varepsilon^{-3})$ iterations using a cubic regularization method.

For specific smooth sets $\mathcal{M}$, Riemannian optimization methods are an efficient way to solve (P), and some have the same worst-case bounds as their unconstrained counterparts [6, 9, 41]. For instance, under a Lipschitz smoothness assumption, a Riemannian trust-region algorithm finds a point which satisfies $\left\|\operatorname{grad}_{\mathcal{M}} f(x)\right\| \leq \varepsilon$ and $\lambda_{\min}(\operatorname{Hess}_{\mathcal{M}} f(x)) \geq -\varepsilon \mathrm{Id}$ in $\mathcal{O}(\varepsilon^{-3})$ iterations.

Riemannian optimization methods are applicable to manifolds $\mathcal{M}$ provided that one is able to compute retractions and generate a feasible sequence of iterates. This is sometimes impossible or too expensive computationally. This prompts the use of infeasible methods to solve (P), which are the focus of this paper.

Several different notions of approximate criticality for (P) are in use in the literature. A systematic review and comparison of existing methods that guarantee second-order criticality for (P) with a complexity analysis is given in [23, Appendix A]. Among those that cover approximate second-order critical points, the rates are either not optimal (worse than $\mathcal{O}(\varepsilon^{-3})$), or they rely on an unusual notion of criticality.

Cartis et al. [13] show optimal complexity rates for finding approximately critical points, which for first- and second-order are respectively $\mathcal{O}(\varepsilon^{-2})$ and $\mathcal{O}(\varepsilon^{-3})$ iterations. Their notion of criticality is unusual but has the advantage of generalizing to optimality beyond second order. They propose a two-phase algorithm that first minimizes infeasibility and then the cost function.

Cifuentes and Moitra [14] adapt the two-phase algorithm from [13] to semidefinite programs that use the Burer–Monteiro factorization. Under Assumptions A1 and A4, along with uniform boundedness and Lipschitz continuity of f, h and their derivatives on $\mathcal{C}$, they show that a second-order critical point can be found in $\mathcal{O}\left(\max\left\{\varepsilon_0^{-2}\varepsilon_1^{-2}, \varepsilon_0^{-3}\varepsilon_2^{-3}\right\}\right)$ iterations. Their definition of critical point is a variant of (1.12) and (1.13) defined below.

Complexity of Augmented Lagrangian Methods A recent point of interest in the literature has been the study of complexity for algorithms that belong to the family of augmented Lagrangian methods (ALM). These methods have always been popular, with good practical results, but worst-case complexity results are lacking [8, 24].

Xie and Wright [40] analyse a proximal ALM, and suggest to solve the subproblems using a Newton-conjugate gradient algorithm from [35]. For this second-order algorithm, they show a total iteration complexity to reach approximate first- and second-order critical points of $\mathcal{O}(\varepsilon^{-11/2})$ and $\mathcal{O}(\varepsilon^{-7})$.

He et al. [26] improved upon the rates of Xie and Wright [40] using a similar method, as they show a total iteration complexity of $\mathcal{O}\left(\varepsilon_1^{-2}\max\left\{\varepsilon_1^{-2}\varepsilon_2, \varepsilon_2^{-3}\right\}\right)$. Both Xie and Wright [40] and He et al. [26] consider that $x \in \mathcal{E}$ is approximately second-order critical if there exists $\lambda \in \mathbb{R}^m$ such that

$$\|h(x)\| \leq \varepsilon_0, \quad \|\nabla_x \mathcal{L}(x, \lambda)\| \leq \varepsilon_1 \tag{1.12}$$

and

$$\left\langle \nabla^2_{xx}\mathcal{L}(x, \lambda)[v], v\right\rangle \geq -\varepsilon_2 \|v\|^2 \text{ for all } v \in \mathcal{E} \text{ such that } \mathrm{D}h(x)[v] = 0. \tag{1.13}$$

Note that the conditions (1.12) and (1.13) are not equivalent to (ε-SOCP). If $x \in \mathcal{E}$ is an (ε-SOCP), then it satisfies (1.12) and (1.13) with multipliers $\lambda(x) \in \mathbb{R}^m$. However, Eqs. (1.12) and (1.13) do not imply (ε-SOCP). For a counter-example, see [23, Appendix A]. The two notions only meet if ε is smaller than some unknown threshold.

Infeasible Optimization Methods for Riemannian Manifolds Riemannian methods can be used to tackle orthogonality constraints. These algorithms are efficient when p is small compared to n, as fast orthogonalization procedures are available. However, when p is large, maintaining orthogonality is often the computational bottleneck [21]. This has prompted the search for retraction-free algorithms to deal with orthogonality constraints. Our use of Fletcher's augmented Lagrangian is partially inspired by Gao et al. [21]. These authors consider the penalty $\hat{g}(x) = \mathcal{L}_\beta(x, \hat{\lambda}(x))$, where $\hat{\lambda}(\cdot)$ is a simplified version of formula (1.7). Ablin and Peyré [1] also present a retraction-free algorithm on the orthogonal group with excellent numerical performance. Schechtman et al. [36] have proposed a recent follow-up work on a first-order method which extends [1] to general manifolds $\mathcal{M}$.

Fletcher's Augmented Lagrangian Around the same year that augmented Lagrangian methods came about, the penalty function we use in this paper—Fletcher's augmented Lagrangian-was introduced in [20]. In the original work, some fundamental properties of the function were established, most notably, connecting critical points of f on $\mathcal{M}$ and critical points of g, see also [7, section 4.3.2]. These properties are covered in Sect. 2, where we extend them to situations with approximate critical points of first- and second-order.

Di Pillo and Grippo [16] Di Pillo [15] present algorithms with local convergence analyses that rely on Fletcher's augmented Lagrangian. Di Pillo and Grippo [17] introduce a 2-norm regularization to compute $\lambda(x)$, which ensures that the multipliers $\lambda(x)$ are well defined even when $\mathrm{D}h(x)$ is singular (outside of the set $\mathcal{D}$ (1.3)). Estrin et al. [18] present a way to compute $g(x)$, $\nabla g(x)$ and approximations of Hessian-vector products $\nabla^2 g(x)v$ that only relies on solving least-square linear systems.

1.4 Contributions

We summarize our contributions in the following list:

- We propose a new definition of approximate criticality for (P), see (ε-FOCP) and (ε-SOCP). These conditions are an extension of Riemannian optimality conditions to points outside the feasible manifold $\mathcal{M}$. We believe that these conditions are more natural geometrically than commonly used conditions for constrained optimization problems. Approximate criticality in our sense implies approximate criticality in the more common sense (with the same ε), but the converse is not true (unless ε is smaller than some unknown threshold).
- We relate the landscape of (P) with the landscape of g (Fletcher's augmented Lagrangian, Eq. 1.2). As far as we know, the existing literature on Fletcher's

Table 1 Summary of related works on complexity for constrained optimization.

Paper	Local rate	Complexity	Target points	Problem class	2nd order
This work	✗*	$\mathcal{O}(\varepsilon^{-2})$ and $\mathcal{O}(\varepsilon^{-3})$	(ε-SOCP)	$\min f(x)$ s.t. $h(x) = 0$	✓
[33]	Quadratic	✗	✗	$\min f(x)$ s.t. $h(x) = 0$	✓
[13]	✗	$\mathcal{O}(\varepsilon^{-2})$ and $\mathcal{O}(\varepsilon^{-3})$	Taylor model	$\min f(x)$ s.t. $h(x) = 0, x \in C$ cvx	✓
[40]	✗	$\mathcal{O}(\varepsilon^{-7})$	(1.12) and (1.13)	$\min f(x)$ s.t. $h(x) = 0$	✓
[14]	✗	$\mathcal{O}(\varepsilon^{-6})$	variant of (1.13)	BM for SDP	✓
[3]	✗	✗	(ε-SOCP) + ineq	$\min f(x)$ s.t. $h(x) = 0, h_2(x) \leq 0$	✓
[39]	Quadratic	✗	(1.11)	$\min f(X)$ s.t. $X \in \mathrm{St}(n, p)$	✓
[24]	✗	$\mathcal{O}(\varepsilon^{-2/(\alpha-1)}), \alpha > 1^{\dagger}$	(ε-FOCP)	$\min f(x)$ s.t. $h(x) = 0, h_2(x) \leq 0$	✗
[21]	Linear	$\mathcal{O}(\varepsilon^{-2})$	(ε-FOCP)	$\min f(X)$ s.t. $X \in \mathrm{St}(n, p)$	✗
[4]	Linear	$\mathcal{O}(\varepsilon^{-4})$	(ε-FOCP)	$\min f(x)$ s.t. $h(x) = 0$	✗
[5]	Linear	✗	(1.10)	$\min f(h(x))$ s.t. $\mathcal{A}(h(x)) = b$, f cvx	✗
[8]	✗	$\mathcal{O}(\log(1/\varepsilon))^{\ddagger}$	(ε-FOCP) + ineq	$\min f(x)$ s.t. $h(x) = 0, h_2(x) \leq 0$	✗

The complexity column gives the total iteration complexity to reach first-order target points and second-order when available. The last column indicates whether second-order critical points are considered.

$*$ The algorithm that we present does not come with a guarantee of local quadratic convergence. However, it is possible to modify it to ensure local quadratic convergence, see Remark 2.1.

† The bound $\mathcal{O}(\varepsilon^{-2/(\alpha-1)})$ in [24] is an outer iteration complexity.

‡ The bound $\mathcal{O}(\log(1/\varepsilon))$ in [8] assumes that the penalty parameters β_k remain bounded as $k \to \infty$. For more details, see Appendix A of the ArXiv version of this paper augmented Lagrangian is limited to asymptotic convergence results. Those rely on the convenient property that exact minimizers of Fletcher's augmented Lagrangian are exact minimizers of (P) under suitable conditions (see Proposition 2.1). In contrast, to obtain complexity bounds, it is necessary to consider approximate minimizers. In Sect. 2, we show that *approximate* first- and second-order critical points of Fletcher's augmented Lagrangian satisfy (ε-FOCP) and (ε-SOCP) for (P), provided that the penalty parameter β is large enough.

- We apply a standard unconstrained minimization algorithm to g-with small modifications to remain in the set $\mathcal{C}$ (1.4)-and leverage our observations from the previous points to deduce guarantees about problem (P). Algorithm 1 finds points which satisfy (ε-FOCP) and (ε-SOCP) for (P) in a worst-case iteration complexity which improves on the state of the art. Our main complexity result is in Theorem 3.5. Informally, it states the following:

Theorem 1.4 (Informal statement) *Under A1, A2, A3, A4, given $\beta > 0$ large enough, Algorithm 1 produces an $(\varepsilon_1, 2\varepsilon_1)$-FOCP of (P) in at most $\mathcal{O}\left(\varepsilon_1^{-2}\right)$ iterations. Algorithm 1 also produces an $(\varepsilon_1, 2\varepsilon_1, \varepsilon_2 + C\varepsilon_1)$-SOCP of (P) in at most $\mathcal{O}\left(\max\{\varepsilon_1^{-2}, \varepsilon_2^{-3}\}\right)$ iterations, where $C \geq 0$ is a constant depending on the constraint function h.*

In Sect. 4 we show how to estimate β at the cost of a log-factor in complexity.

- Algorithm 1 is the first augmented Lagrangian method to find approximate second-order critical points in a total iteration complexity of $\mathcal{O}(\varepsilon^{-3})$. Beyond augmented Lagrangian methods, the only other method that we are aware of which achieves a similar complexity is the two-phase method in [13]; however, this method achieves a markedly different notion of criticality which makes the comparison delicate. Table 1 summarizes some of the main results.

Admittedly, the class of problems to which our theoretical algorithm applies is somewhat limited. These limitations should be compared to those of other algorithm with comparable theoretical guarantees, which also feature restrictive assumptions. Other works face similar limitations in terms of initialization, smoothness and compactness assumptions. When other works are able to relax some of these assumptions, it typically comes with a worsening of the iteration complexity.

2 Properties of Fletcher's Augmented Lagrangian

We now cover properties of the function g, Fletcher's augmented Lagrangian (Eq. 1.2). In this section, we recall an original result from [7] which establishes conditions under which the critical points and minimizers of g and (P) are equivalent. The core of this section then establishes extensions of this result to the case of approximate critical points. That is, we show that approximate first- and second-order critical points of g are also approximately critical for (P) in the sense of (ε-FOCP) and (ε-SOCP).

For problem (P), the Lagrangian $\mathcal{L}(x, \lambda) \colon \mathcal{E} \times \mathbb{R}^m \to \mathbb{R}$ is defined as

$$\mathcal{L}(x, \lambda) = f(x) - \langle \lambda, h(x) \rangle ,$$

where $\lambda \in \mathbb{R}^m$ is called the vector of multipliers. The augmented Lagrangian $\mathcal{L}_\beta : \mathcal{E} \times \mathbb{R}^m \to \mathbb{R}$ for some penalty parameter $\beta \geq 0$ is:

$$\mathcal{L}_\beta(x, \lambda) = f(x) - \langle \lambda, h(x) \rangle + \beta \|h(x)\|^2. \tag{2.1}$$

This penalty function has given rise to a number of popular methods for constrained optimization [7]. Fletcher [20] proposed a variant, which we denote by g (already shown in Eq. 1.2):

$$g(x) = \mathcal{L}_\beta(x, \lambda(x)), \tag{2.2}$$

where $\lambda(x)$ is defined in (1.7). We note that the set $\mathcal{D}$ (Eq. 1.3) is open, and it is easy to verify that $\lambda(\cdot)$ is C^∞ on that set. We also note that, under A1, the set $\mathcal{C}$ is included in $\mathcal{D}$. Therefore, g is also smooth on $\mathcal{C}$. Fletcher's augmented Lagrangian is a smooth penalty which depends only on x, the primal variable. The multipliers are computed as a function of x. We define $C_\lambda(x)$ as the operator norm of the differential of $\lambda(\cdot)$. Since $\mathcal{C}$ is assumed compact and $\lambda(\cdot)$ is smooth, this quantity is bounded.

Definition 2.1 Under A1, for any $x \in \mathcal{C}$, we define the quantity

$$C_\lambda(x) := \|\mathrm{D}\lambda(x)\|_{\mathrm{op}} = \sigma_1\left(\mathrm{D}\lambda(x)\right).$$

Additionally, under A2, we define the constant

$$\overline{C_\lambda} := \max_{x \in \mathcal{C}} \|\mathrm{D}\lambda(x)\|_{\mathrm{op}} < \infty.$$

Definition 2.2 Under A1, for $x \in \mathcal{C}$, we define the following quantities

$$\beta_1(x) = \frac{\sigma_1(\mathrm{D}h(x)) C_\lambda(x)}{2\sigma_{\min}^2(\mathrm{D}h(x))} \tag{2.3}$$

$$\beta_2(x) = \frac{C_\lambda(x)}{\sigma_{\min}(\mathrm{D}h(x))} \tag{2.4}$$

$$\beta_3(x) = \frac{1}{\sigma_{\min}(\mathrm{D}h(x))}. \tag{2.5}$$

Additionally, under A2, we define the constants $\bar{\beta_i} = \max_{x \in \mathcal{C}} \beta_i(x)$ for $i = 1, 2, 3$.

The following classical result connects first-order critical points and minimizers of g and (P).

Proposition 2.1 ([7], *Prop. 4.22) Let $g(x) = \mathcal{L}_\beta(x, \lambda(x))$ be Fletcher's augmented Lagrangian and assume $\mathcal{M} \subset \mathcal{D}$, where $\mathcal{D} = \{x \in \mathcal{E} : \mathrm{rank}(\mathrm{D}h(x)) = m\}$ and $\mathcal{M} = \{x \in \mathcal{E} : h(x) = 0\}$.*

1. *For any β, if x is a first-order critical point of (P), then x is a first-order critical point of g.*

2. *Let $x \in \mathcal{D}$ and $\beta > \beta_1(x)$. If x is a first-order critical point of g, then x is a first-order critical point of (P).*
3. *Let x be a first-order critical point of (P) and let K be a compact set. Assume x is the unique global minimum of f over $\mathcal{M} \cap K$ and that x is in the interior of K. Then, there exists β large enough such that x is the unique global minimum of g over K.*
4. *Let $x \in \mathcal{D}$ and $\beta > \beta_1(x)$. If x is a local minimum of g, then x is a local minimum of (P).*

The previous shows that minimizing the function g inside $\mathcal{D}$ provides a way to find minimizers of (P). However, in practice, algorithms can only find approximate first- and second-order critical points in finite time. With the above proposition, one is left wondering whether such approximate points for g correspond to similarly approximate critical points for (P). The remainder of this section provides such guarantees.

2.1 Approximate First-Order Criticality

In this section, we show that if $\nabla g(x)$ is small at some $x \in \mathcal{C}$, the point x is approximately first-order critical for (P) in the sense of (ε-FOCP). We begin with a straightforward computation of the gradient of g. The gradient of the augmented Lagrangian $\mathcal{L}_\beta$ with respect to x is given by

$$\begin{aligned}\nabla_x \mathcal{L}_\beta(x, \lambda) &= \nabla f(x) - \mathrm{D}h(x)^*[\lambda] + 2\beta \mathrm{D}h(x)^*[h(x)] \\ &= \nabla f(x) - \mathrm{D}h(x)^*[\lambda - 2\beta h(x)]. \end{aligned} \tag{2.6}$$

Owing to (1.8), we make the following central observation: the gradient of $\mathcal{L}_\beta$ with respect to its first argument, when evaluated at $(x, \lambda(x))$, splits into orthogonal components; one component in the tangent space $\mathrm{T}_x\mathcal{M}_x$, and one component in the normal space to $\mathcal{M}_x$ at x:

$$\nabla_x \mathcal{L}_\beta(x, \lambda(x)) = \mathrm{grad}_{\mathcal{M}_x} f(x) + 2\beta \mathrm{D}h(x)^*[h(x)]. \tag{2.7}$$

Owing to orthogonality, $\nabla_x \mathcal{L}_\beta(x, \lambda(x))$ is small if and only if the two terms on the right are small. It takes an easy computation to check that for all $x \in \mathcal{D}$ we have

$$\begin{aligned}\mathrm{D}g(x)[v] &= \mathrm{D}f(x)[v] - \langle \mathrm{D}\lambda(x)[v], h(x)\rangle - \langle \lambda(x), \mathrm{D}h(x)[v]\rangle + 2\beta \langle h(x), \mathrm{D}h(x)[v]\rangle \\ &= \langle \nabla f(x), v\rangle - \left\langle \mathrm{D}h(x)^*[\lambda(x) - 2\beta h(x)], v\right\rangle - \left\langle \mathrm{D}\lambda(x)^*[h(x)], v\right\rangle \\ &= \left\langle \nabla_x \mathcal{L}_\beta(x, \lambda(x)), v\right\rangle - \left\langle \mathrm{D}\lambda(x)^*[h(x)], v\right\rangle.\end{aligned}$$

Thus, for all $x \in \mathcal{D}$,

$$\begin{aligned}\nabla g(x) &= \nabla_x \mathcal{L}_\beta(x, \lambda(x)) - \mathrm{D}\lambda(x)^*[h(x)] \\ &= \mathrm{grad}_{\mathcal{M}_x} f(x) + 2\beta \mathrm{D}h(x)^*[h(x)] - \mathrm{D}\lambda(x)^*[h(x)].\end{aligned} \tag{2.8}$$

Therefore, for $x \in \mathcal{M}$, $\nabla g(x) = \mathrm{grad}_{\mathcal{M}} f(x)$. Consequently, for any value of β, if x satisfies the constraints $h(x) = 0$, that is, if x is on the manifold $\mathcal{M}$, then x is first-order critical for f on $\mathcal{M}$ (Eq. 1.10) if and only if $\nabla g(x) = 0$. We now add to this with a claim about approximate first-order critical points of g.

Proposition 2.2 *Under A1, take $\varepsilon_1 \geq 0$ and $x \in \mathcal{C}$ with $\beta > \max\{\beta_2(x), \beta_3(x)\}$. If $\|\nabla g(x)\| \leq \varepsilon_1$, then x is an $(\varepsilon_1, 2\varepsilon_1)-$approximate first-order critical point of (P) (see (ε-FOCP)) as*

$$\|h(x)\| \leq \frac{\varepsilon_1}{\beta \sigma_{\min}(\mathrm{D}h(x))} \leq \varepsilon_1 \quad \textit{and} \quad \left\|\mathrm{grad}_{\mathcal{M}_x} f(x)\right\| \\ \leq \left(1 + \frac{C_\lambda(x)}{\beta \sigma_{\min}(\mathrm{D}h(x))}\right) \varepsilon_1 \leq 2\varepsilon_1. \tag{2.9}$$

Proof We remember from (2.8) that

$$\nabla g(x) = \mathrm{grad}_{\mathcal{M}_x} f(x) + 2\beta \mathrm{D}h(x)^*[h(x)] - \mathrm{D}\lambda(x)^*[h(x)] \tag{2.10}$$

$$= \mathrm{grad}_{\mathcal{M}_x} f(x) - \mathrm{Proj}_x\left(\mathrm{D}\lambda(x)^*[h(x)]\right) + 2\beta \mathrm{D}h(x)^*[h(x)] \\ - \mathrm{Proj}_x^{\perp}\left(\mathrm{D}\lambda(x)^*[h(x)]\right) \tag{2.11}$$

where $\mathrm{Proj}_x^{\perp} = \mathrm{Id} - \mathrm{Proj}_x$, is the orthogonal projection on $\mathrm{N}_x\mathcal{M}_x = (\mathrm{T}_x\mathcal{M}_x)^{\perp}$, the normal space to $\mathcal{M}_x$ at x. We have decomposed the right-hand side in two tangent and two normal terms with respect to the manifold $\mathcal{M}_x$. By orthogonality, $\|\nabla g(x)\| \leq \varepsilon_1$ implies that both the tangent and normal components have norm smaller than ε_1. For the normal terms this yields,

$$\left\|2\beta \mathrm{D}h(x)^*[h(x)] - \mathrm{Proj}_x^{\perp}\left(\mathrm{D}\lambda(x)^*[h(x)]\right)\right\| \\ = \left\|\left(2\beta \mathrm{D}h(x)^* - \mathrm{Proj}_x^{\perp}\left(\mathrm{D}\lambda(x)^*\right)\right)[h(x)]\right\| \leq \varepsilon_1. \tag{2.12}$$

Note that $\mathrm{D}h(x)^*$ is nonsingular since $x \in \mathcal{C}$. We show that β is large enough so that the operator $\left(2\beta \mathrm{D}h(x)^* - \mathrm{Proj}_x^{\perp}\left(\mathrm{D}\lambda(x)^*\right)\right)$ is nonsingular. We use Weyl's inequality to control singular values (Horn and Johnson [27, Theorem 3.3.16 (c)]), which states that for two linear operators $A, B \colon \mathcal{E}_1 \to \mathcal{E}_2$, it holds that $\sigma_q(A - B) \geq \sigma_q(A) - \sigma_1(B)$ with $q \leq \min(m, n)$. This allows to write

$$\sigma_{\min}\left(2\beta \mathrm{D}h(x)^* - \mathrm{Proj}_x^{\perp}\left(\mathrm{D}\lambda(x)^*\right)\right) \\ \geq \sigma_{\min}\left(2\beta \mathrm{D}h(x)^*\right) - \sigma_{\max}\left(\mathrm{Proj}_x^{\perp}\left(\mathrm{D}\lambda(x)^*\right)\right). \tag{2.13}$$

The assumption on β then provides

$$\sigma_{\min}\left(2\beta \mathrm{D}h(x)^*\right) - \sigma_{\max}\left(\mathrm{Proj}_x^{\perp}\left(\mathrm{D}\lambda(x)^*\right)\right) \geq 2\beta \sigma_{\min}(\mathrm{D}h(x)) - C_\lambda(x) \tag{2.14}$$

$$> \beta \sigma_{\min}(\mathrm{D}h(x)) > 1. \tag{2.15}$$

We inject this into (2.12) to find:

$$\|h(x)\| \leq \frac{\varepsilon_1}{\sigma_{\min}\left(2\beta \mathrm{D}h(x)^* - \mathrm{Proj}_x^{\perp}\left(\mathrm{D}\lambda(x)^*\right)\right)} \tag{2.16}$$

$$\leq \frac{\varepsilon_1}{\beta\sigma_{\min}(\mathrm{D}h(x))} \leq \varepsilon_1. \tag{2.17}$$

Now we use the tangent terms:

$$\begin{aligned}
\varepsilon_1 &\geq \left\|\mathrm{grad}_{\mathcal{M}_x} f(x) - \mathrm{Proj}_x\left(\mathrm{D}\lambda(x)^*[h(x)]\right)\right\| \\
&\geq \left\|\mathrm{grad}_{\mathcal{M}_x} f(x)\right\| - \left\|\mathrm{Proj}_x\left(\mathrm{D}\lambda(x)^*[h(x)]\right)\right\| \\
&\geq \left\|\mathrm{grad}_{\mathcal{M}_x} f(x)\right\| - \left\|\mathrm{D}\lambda(x)^*[h(x)]\right\| \\
&\geq \left\|\mathrm{grad}_{\mathcal{M}_x} f(x)\right\| - C_\lambda(x)\,\|h(x)\| \\
&\geq \left\|\mathrm{grad}_{\mathcal{M}_x} f(x)\right\| - C_\lambda(x)\frac{\varepsilon_1}{\beta\sigma_{\min}(\mathrm{D}h(x))}.
\end{aligned} \tag{2.18}$$

This allows to conclude $\left\|\mathrm{grad}_{\mathcal{M}_x} f(x)\right\| \leq \varepsilon_1 + \frac{C_\lambda(x)}{\beta\sigma_{\min}(\mathrm{D}h(x))}\varepsilon_1 \leq 2\varepsilon_1.$ □

Corollary 2.3 *Under A1 and A2, take $\varepsilon_1 \geq 0$. Let β satisfies the global bounds*

$$\beta > \bar{\beta}_2 \quad \textit{and} \quad \beta > \bar{\beta}_3, \tag{2.19}$$

where $\bar{\beta}_2, \bar{\beta}_3$ are introduced in Definition 2.2. Any $x \in \mathcal{C}$ such that $\|\nabla g(x)\| \leq \varepsilon_1$ is an $(\varepsilon_1, 2\varepsilon_1)-$approximate first-order critical point of (P) (see (ε-FOCP)) as

$$\|h(x)\| \leq \frac{\varepsilon_1}{\beta\underline{\sigma}} \leq \varepsilon_1 \qquad \textit{and} \qquad \left\|\mathrm{grad}_{\mathcal{M}_x} f(x)\right\| \leq \left(1 + \frac{\overline{C_\lambda}}{\beta\underline{\sigma}}\right)\varepsilon_1 \leq 2\varepsilon_1, \tag{2.20}$$

with $\underline{\sigma} \leq \min_{x\in\mathcal{C}} \sigma_{\min}\left(\mathrm{D}h(x)\right)$ defined in A1.

2.2 Approximate Second-Order Criticality

We now turn our attention to approximate second-order critical points of Fletcher's augmented Lagrangian. Similarly to first-order criticality, we investigate connections with (ε-SOCP) points for (P). Specifically, we extend the observation that *strict* second-order critical points of (P) and g match, provided that β is large enough. The non-strict case is less clear: see below.

The Hessian of g is obtained by taking a directional derivative of (2.8). For any $\dot{x} \in \mathcal{E}$,

$$\begin{aligned}
\nabla^2 g(x)[\dot{x}] = \nabla^2 f(x)[\dot{x}] \\
&- \left(\mathrm{D}\left(x \mapsto \mathrm{D}\lambda(x)^*\right)(x)[\dot{x}]\right)[h(x)] \\
&- \mathrm{D}\lambda(x)^*[\mathrm{D}h(x)[\dot{x}]]
\end{aligned}$$

$$
\begin{aligned}
&- \big(\mathrm{D}\big(x \mapsto \mathrm{D}h(x)^*\big)(x)[\dot{x}]\big)\,[\lambda(x) - 2\beta h(x)] \\
&- \mathrm{D}h(x)^*[\mathrm{D}\lambda(x)[\dot{x}] - 2\beta \mathrm{D}h(x)[\dot{x}]]\,. \qquad (2.21)
\end{aligned}
$$

We begin with a statement about feasible points which connects the Hessian of g and the Riemannian Hessian of f on $\mathcal{M}$.

Proposition 2.4 *For all $x \in \mathcal{M}$, with Proj_x the orthogonal projector from $\mathcal{E}$ to $\mathrm{T}_x\mathcal{M}$, we have*

$$
\mathrm{Hess}_{\mathcal{M}} f(x) = \left(\mathrm{Proj}_x \circ \nabla^2 g(x) \circ \mathrm{Proj}_x\right)\Big|_{\mathrm{T}_x\mathcal{M}}. \qquad (2.22)
$$

Therefore, if $\nabla^2 g(x) \succeq -\varepsilon_2 \mathrm{Id}$, *then* $\mathrm{Hess}_{\mathcal{M}} f(x) \succeq -\varepsilon_2 \mathrm{Id}_{\mathrm{T}_x\mathcal{M}}$. *If* $\nabla^2 g(x) \succ 0$, *then* $\mathrm{Hess}_{\mathcal{M}} f(x) \succ 0$.

Proof We show that if $h(x) = 0$, then (2.22) holds. Take $\dot{x} \in \mathcal{E}$ and plug $h(x) = 0$ into Eq. (2.21). This gives

$$
\begin{aligned}
\nabla^2 g(x)[\dot{x}] = \nabla^2 f(x)[\dot{x}] - \sum_{i=1}^{m} \lambda_i(x) \nabla^2 h_i(x)[\dot{x}] \\
+ 2\beta \mathrm{D}h(x)^*[\mathrm{D}h(x)[\dot{x}]] \\
- \mathrm{D}\lambda(x)^*[\mathrm{D}h(x)[\dot{x}]] - \mathrm{D}h(x)^*[\mathrm{D}\lambda(x)[\dot{x}]]\,. \qquad (2.23)
\end{aligned}
$$

If, in addition, $\dot{x} \in \ker \mathrm{D}h(x)$, then

$$
\left\langle \dot{x}, \nabla^2 g(x)[\dot{x}] \right\rangle = \left\langle \dot{x}, \nabla^2 f(x)[\dot{x}] \right\rangle - \sum_{i=1}^{m} \lambda_i(x) \left\langle \dot{x}, \nabla^2 h_i(x)[\dot{x}] \right\rangle.
$$

Since $\ker \mathrm{D}h(x) = \mathrm{T}_x\mathcal{M}$, we conclude from Eq. (1.9) that, restricted to $\mathrm{T}_x\mathcal{M}$,

$$
\begin{aligned}
\mathrm{Proj}_x \circ \nabla^2 g(x) \circ \mathrm{Proj}_x &= \mathrm{Proj}_x \circ \left(\nabla^2 f(x) - \sum_{i=1}^{m} \lambda_i(x) \nabla^2 h_i(x)\right) \circ \mathrm{Proj}_x \\
&= \mathrm{Hess}_{\mathcal{M}} f(x).
\end{aligned}
$$

□

In particular, for $\varepsilon_2 = 0$, the above result tells us that, irrespective of $\beta \geq 0$, if $x \in \mathcal{M}$ satisfies $\nabla g(x) = 0$ and $\nabla^2 g(x) \succeq 0$, the point x is second-order critical for f on $\mathcal{M}$ (Eq. 1.11). To our knowledge, there is no evidence that the converse is true, namely, we do not know whether at a point $x \in \mathcal{M}$ that satisfies (1.11) there exists a β large enough such that $\nabla^2 g(x) \succeq 0$. Fletcher [20] showed that the converse holds for positive definite Hessians.

Proposition 2.5 *(Fletcher [20]) If $x \in \mathcal{M}$ is a local minimizer of (P) with* $\mathrm{Hess}_{\mathcal{M}} f(x) \succ 0$, *there exists β large enough such that* $\nabla^2 g(x) \succ 0$.

Remark 2.1 (*Local quadratic convergence*) Assume that A1, A2 and A4 hold. For β large enough, it is possible to apply Newton's method to Fletcher's augmented Lagrangian to achieve a local quadratic convergence rate towards isolated minimizers of (P). Let $x^* \in \mathcal{M}$ be a strict second-order critical point for (P) satisfying $\mathrm{grad}_{\mathcal{M}} f(x^*) = 0$ and $\mathrm{Hess}_{\mathcal{M}} f(x^*) \succ 0$. Provided β is large enough, x^* satisfies $\nabla g(x^*) = 0$ and $\nabla^2 g(x^*) \succ 0$ (Propositions 2.1 and 2.5). Take $x_0 \in \mathcal{C}$ close enough to x^*, the classical Newton method applied to the function g produces a sequence which converges towards x^* at a quadratic rate, as is discussed in [18].

Proposition 2.4 can be generalized to infeasible points in $\mathcal{C}$ using an upper bound on the gradient norm of g.

Proposition 2.6 *Under* A1*, take $x \in \mathcal{C}$ with $\beta > \max\{\beta_2(x), \beta_3(x)\}$. Assume $\|\nabla g(x)\| \leq \varepsilon_1$ so that Proposition* 2.2 *applies at x. If $\nabla^2 g(x) \succeq -\varepsilon_2 \mathrm{Id}$, then x is an $(\varepsilon_1, 2\varepsilon_1, \varepsilon_2 + C(x)\varepsilon_1)$−approximate second-order critical point of* (P) *(see (*ε-SOCP*)) as*

$$\mathrm{Hess}_{\mathcal{M}_x} f(x) \succeq -(\varepsilon_2 + C(x)\varepsilon_1)\mathrm{Id}, \tag{2.24}$$

where $C(x) = 2\,\|(\mathrm{D}(x \mapsto \mathrm{D}h(x)^*)(x))\|_{\mathrm{op}} / \sigma_{\min}(\mathrm{D}h(x)) + \|(\mathrm{D}(x \mapsto \mathrm{D}\lambda(x)^*)(x))\|_{\mathrm{op}}$.

Proof Since $x \in \mathcal{C}$, for any $\dot{x} \in \mathrm{T}_x\mathcal{M}_x$, Eq. (1.9) gives the Riemannian Hessian of f at x and yields

$$\langle \dot{x}, \mathrm{Hess}_{\mathcal{M}_x} f(x)[\dot{x}] \rangle = \left\langle \dot{x}, \nabla^2 f(x)[\dot{x}] - \sum_{i=1}^{m} \lambda_i(x) \nabla^2 h_i(x)[\dot{x}] \right\rangle. \tag{2.25}$$

By assumption, for any $\dot{x} \in \mathcal{E}$,

$$\langle \dot{x}, \nabla^2 g(x)[\dot{x}] \rangle \geq -\varepsilon_2 \|\dot{x}\|^2. \tag{2.26}$$

In Eq. (2.21), take $\dot{x} \in \mathrm{T}_x\mathcal{M}_x = \ker(\mathrm{D}h(x))$ and remember that the span of $\mathrm{D}h(x)^*$ is orthogonal to $\mathrm{T}_x\mathcal{M}_x$. This gives

$$\begin{aligned}\left(\mathrm{Proj}_x \circ \nabla^2 g(x) \circ \mathrm{Proj}_x\right)[\dot{x}] = \mathrm{Proj}_x \circ \Big(&\nabla^2 f(x)[\dot{x}] \\ &- \left(\mathrm{D}\big(x \mapsto \mathrm{D}h(x)^*\big)(x)[\dot{x}]\right)[\lambda(x) - 2\beta h(x)] \\ &- \left(\mathrm{D}\big(x \mapsto \mathrm{D}\lambda(x)^*\big)(x)[\dot{x}]\right)[h(x)]\Big).\end{aligned} \tag{2.27}$$

For clarity, we write $F_h(x) = \mathrm{D}(x \mapsto \mathrm{D}h(x)^*)(x)$ and $F_\lambda(x) = \mathrm{D}(x \mapsto \mathrm{D}\lambda(x)^*)(x)$. We compute the derivative

$$-\left(F_h(x)[\dot{x}]\right)[\lambda(x) - 2\beta h(x)] = -\sum_{i=1}^{m} (\lambda_i(x) - 2\beta h_i(x)) \nabla^2 h_i(x)[\dot{x}], \tag{2.28}$$

which gives

$$\begin{aligned}\left\langle \dot{x}, \mathrm{Proj}_x \circ \nabla^2 g(x) \circ \mathrm{Proj}_x[\dot{x}] \right\rangle &= \left\langle \dot{x}, \nabla^2 f(x)[\dot{x}] - \sum_{i=1}^{m} \lambda_i(x) \nabla^2 h_i(x)[\dot{x}] \right\rangle \\ &\quad + \langle \dot{x}, 2\beta\, (F_h(x)[\dot{x}])\, [h(x)] \rangle \\ &\quad - \langle \dot{x}, (F_\lambda(x)[\dot{x}])\, [h(x)] \rangle \\ &\geq -\varepsilon_2 \left\| \dot{x} \right\|^2 . \end{aligned} \tag{2.29}$$

The formula for $\mathrm{Hess}_{\mathcal{M}_x} f(x)$ has appeared on the right-hand side. Using $\|h(x)\| \leq \dfrac{\varepsilon_1}{\beta \sigma_{\min}(\mathrm{D}h(x))} \leq \varepsilon_1$ from Proposition 2.2, we conclude with

$$\begin{aligned}\left\langle \dot{x}, \nabla^2 f(x)[\dot{x}] - \sum_{i=1}^{m} \lambda_i(x) \nabla^2 h_i(x)[\dot{x}] \right\rangle &\geq -2\beta \langle \dot{x}, (F_h(x)[\dot{x}])\, [h(x)] \rangle \\ &\quad + \langle \dot{x}, (F_\lambda(x)[\dot{x}])\, [h(x)] \rangle - \varepsilon_2 \left\| \dot{x} \right\|^2 \\ &\geq -2\beta \left\| F_h(x) \right\|_{\mathrm{op}} \left\| \dot{x} \right\|^2 \left\| h(x) \right\| \\ &\quad - \left\| F_\lambda(x) \right\|_{\mathrm{op}} \left\| \dot{x} \right\|^2 \left\| h(x) \right\| - \varepsilon_2 \left\| \dot{x} \right\|^2 \\ &\geq -2\beta \left\| F_h(x) \right\|_{\mathrm{op}} \left\| \dot{x} \right\|^2 \frac{\varepsilon_1}{\beta \sigma_{\min}(\mathrm{D}h(x))} \\ &\quad - \left\| F_\lambda(x) \right\|_{\mathrm{op}} \left\| \dot{x} \right\|^2 \varepsilon_1 - \varepsilon_2 \left\| \dot{x} \right\|^2 \\ &\geq -\varepsilon_2 \left\| \dot{x} \right\|^2 - \big(2 \left\| F_h(x) \right\|_{\mathrm{op}} / \sigma_{\min}(\mathrm{D}h(x)) \\ &\quad + \left\| F_\lambda(x) \right\|_{\mathrm{op}} \big) \varepsilon_1 \left\| \dot{x} \right\|^2 . \end{aligned} \tag{2.30}$$

□

Corollary 2.7 *Under A1 and A2, let $\beta > \max\left\{\bar{\beta}_2, \bar{\beta}_3\right\}$. Take $x \in \mathcal{C}$ with $\|\nabla g(x)\| \leq \varepsilon_1$ so that Corollary 2.3 applies. If $\nabla^2 g(x) \succeq -\varepsilon_2 Id$, then x is an $(\varepsilon_1, 2\varepsilon_1, \varepsilon_2 + C\varepsilon_1)-$approximate second-order critical point of (P) (see (ε-SOCP)) as*

$$\mathrm{Hess}_{\mathcal{M}_x} f(x) \succeq -(\varepsilon_2 + C\varepsilon_1)\mathrm{Id}, \tag{2.31}$$

where $C = \max_{x \in \mathcal{C}} 2 \left\| (\mathrm{D}(x \mapsto \mathrm{D}h(x)^*)(x)) \right\|_{\mathrm{op}} / \underline{\sigma} + \left\| (\mathrm{D}(x \mapsto \mathrm{D}\lambda(x)^*)(x)) \right\|_{\mathrm{op}}$.

2.3 Property of the Region $\mathcal{C}$

The algorithms we design and analyse in later sections are initialized in some connected component of $\mathcal{C} = \{x \in \mathcal{E} : \|h(x)\| \leq R\}$, with R as in A1, and produce iterates which remain in this same connected component. Since $\mathcal{C}$ may in general have more than one such component, and since we hope in particular that our iterates converge to a feasible point, that is, to a point in $\mathcal{M} = \{x \in \mathcal{E} : h(x) = 0\}$, it is natural to wonder whether each connected component of $\mathcal{C}$ intersects with $\mathcal{M}$. That is indeed the case. We prove the following result in Appendix A.

Proposition 2.8 *Under A1, every connected component of $\mathcal{C}$ contains a point $\bar{z} \in \mathcal{E}$ such that $h(\bar{z}) = 0$.*

The proof relies on the escape lemma from differential equations [28, Theorem A.42] with a classical Polyak–Łojasiewicz (PŁ) condition along gradient flows. The latter part is similar in nature to Theorem 9 in [32], but the assumptions of the latter are too strong for our purpose. The cited theorem requires (among other things) that the PŁ condition hold in a ball centered around x_0 whose size may be large depending on problem constants, while in our case, PŁ only holds in $\mathcal{C}$.

3 Gradient-Eigenstep Algorithm

In light of the properties of the approximate minimizers of Fletcher's augmented Lagrangian established in the previous section, it would be natural to use an off-the-shelf algorithm for unconstrained minimization on the function g. However, we need to ensure that the iterates remain in the set $\mathcal{C}$, which is not automatic. To this end, we present in this section an optimization algorithm to minimize g that is designed to remain in $\mathcal{C}$, the region of interest where $\lambda(x)$ is well defined. The algorithm alternates between gradient steps (first-order) and eigensteps (second-order) to reach approximate second-order critical points of g, as described in [37, Section 3.6]. If the gradient norm of g is large, a gradient step on g is used. If the gradient of g is below a tolerance, the algorithm follows a direction of negative curvature of the Hessian of g. Gradient steps and eigensteps must fulfil two purposes: they must guarantee a sufficient decrease of the penalty g and also ensure that the next iterate remains inside $\mathcal{C}$. This is detailed in Algorithm 1. Given values $\varepsilon_1 > 0$, $\varepsilon_2 > 0$, the algorithm returns a point which satisfies $\|\nabla g(x)\| \leq \varepsilon_1$ and $\lambda_{\min}\left(\nabla^2 g(x)\right) \geq -\varepsilon_2$. This ensures that x is an $(\varepsilon_1, \varepsilon_2 + C(x)\varepsilon_1)$-SOCP of (P) according to Proposition 2.6.

Whenever $\|\nabla g(x)\| > \varepsilon_1$, a gradient step is used and we require that the step-length α satisfies a classical Armijo sufficient decrease condition:

$$g(x) - g(x - \alpha \nabla g(x)) \geq c_1 \alpha \|\nabla g(x)\|^2, \tag{3.1}$$

for some $0 < c_1 < 1$. The backtracking procedure for gradient steps is presented in Algorithm 2. This is a classical backtracking modified to additionally ensure that the iterates stay in $\mathcal{C}$, which is always possible for small enough steps, as we show in Proposition 3.1.

Given $x \in \mathcal{C}$ with $\|\nabla g(x)\| \leq \varepsilon_1$ and $\lambda_{\min}\left(\nabla^2 g(x)\right) < -\varepsilon_2$, a second-order step must be applied. We compute a unit-norm vector $d \in \mathcal{E}$ such that $\langle d, \nabla^2 g(x)[d]\rangle < -\varepsilon_2 \|d\|^2$. To ensure sufficient decrease, we wish to find $\alpha > 0$ such that

$$g(x) - g(x + \alpha d) \geq -c_2 \alpha^2 \langle d, \nabla^2 g(x)[d]\rangle, \tag{3.2}$$

for some $0 < c_2 < 1/2$. In Algorithm 3, we detail the backtracking used for second-order steps. It is designed to ensure that (3.2) holds and additionally that the steps are small enough for the iterates to remain in $\mathcal{C}$, which is possible as we show in Proposition 3.3.

We define some bounds on the derivatives of g, which are finite due to the smoothness of g in $\mathcal{C}$ and boundedness of $\mathcal{C}$ (A2).

Definition 3.1 Under A1 and A2, define the constants

$$L_g = \max_{x\in\mathcal{C}} \left\|\nabla^2 g(x)\right\|_{\mathrm{op}} \qquad \text{and} \qquad M_g = \max_{x\in\mathcal{C}} \left\|\nabla^3 g(x)\right\|_{\mathrm{op}}. \tag{3.3}$$

3.1 Algorithm

We define Algorithm 1, a procedure which combines first- and second-order steps to minimize g up to approximate second-order criticality if $\varepsilon_2 < \infty$. Setting $\varepsilon_2 = \infty$ gives a first-order version of the algorithm. To run Algorithm 1, we assume that the value of the penalty parameter β does not change and is large enough in the following sense.

A 5 Under A1 and A2, β is chosen such that $\beta > \overline{\beta}$ with

$$\overline{\beta} := \max\left\{\overline{\beta}_1, \overline{\beta}_2, \overline{\beta}_3\right\}, \tag{3.4}$$

where $\overline{\beta}_i$ for $i = 1, 2, 3$ are defined in Definition 2.2.

In Sect. 4, we show how this assumption can be removed, using an adaptive scheme for β.

Algorithm 1 Gradient-Eigenstep

```
Given: Functions $f$ and $h$, $x_0 \in \mathcal{C}$, $\beta > 0$, $0 \le \varepsilon_1 \le R/2$ and $\varepsilon_2 \ge 0$.
Set $k \leftarrow 0$
while no optional stopping criterion triggers do
    if $\|\nabla g(x_k)\| > \varepsilon_1$ then
        $x_{k+1} = x_k - t\nabla g(x_k)$ with $t$ given by Algorithm 2
    else if $\varepsilon_2 < \infty$ then
        if $\lambda_{\min}(\nabla^2 g(x_k)) < -\varepsilon_2$ then
            Find $d \in \mathcal{E}$ such that $\langle d, \nabla^2 g(x_k)[d]\rangle < -\varepsilon_2 \|d\|^2$, $\langle d, \nabla g(x_k)\rangle \le 0$ and $\|d\| = 1$.
            $x_{k+1} = x_k + td$ where $t$ is given by Algorithm 3.
        else
            return $x_k$                                ▷ $\|\nabla g(x_k)\| \le \varepsilon_1$ and $\nabla^2 g(x_k) \succeq -\varepsilon_2 \mathrm{Id}$
        end if
    else
        return $x_k$                                    ▷ $\|\nabla g(x_k)\| \le \varepsilon_1$
    end if
    $k \leftarrow k + 1$
end while
```

3.2 First-Order Steps

We show that small enough gradient steps remain in the set $\mathcal{C}$ (defined in A1).

Algorithm 2 Gradient step backtracking, modified to stay in $\mathcal{C}$

```
Given: $x \in \mathcal{C}$, $\alpha_{01} > 0$, $0 < c_1 < 1$, $0 < \tau_1 < 1$.
Set $\alpha \leftarrow \alpha_{01}$
while true do
    if $g(x) - g(x - \alpha \nabla g(x)) \geq c_1 \alpha \|\nabla g(x)\|^2$ and $x - \alpha \nabla g(x) \in \mathcal{C}$ then
        return $\alpha$
    else
        $\alpha \leftarrow \tau_1 \alpha$
    end if
end while
```

Algorithm 3 Eigenstep backtracking, modified to stay in $\mathcal{C}$

```
Given: $x \in \mathcal{C}$, unit-norm $d \in \mathcal{E}$, $\alpha_{02} > 0$, $0 < c_2 < 1/2$, $0 < \tau_2 < 1$.
Set $\alpha \leftarrow \alpha_{02}$
while true do
    if $g(x) - g(x + \alpha d) \geq -c_2 \alpha^2 \langle d, \nabla^2 g(x)[d] \rangle$ and $x + \alpha d \in \mathcal{C}$ then
        return $\alpha$
    else
        $\alpha \leftarrow \tau_2 \alpha$
    end if
end while
```

Proposition 3.1 *Assume A1 holds with constant R and A3 holds with constant C_h. Then, for all $x \in \mathcal{C}$, if $\beta > \beta_1(x)$, it holds that $x - t\nabla g(x)$ is in $\mathcal{C}$ for all t in the interval $[0, t_1(x)]$ where $t_1(x)$ is defined by*

$$t_1(x) := \min\left(\sqrt{\frac{R}{2C_h}} \frac{1}{\|\nabla g(x)\|}, \frac{(2\beta\sigma_{\min}(\mathrm{D}h(x))^2 - \sigma_1(\mathrm{D}h(x))C_\lambda(x))R}{2C_h \|\nabla g(x)\|^2}, \frac{1}{2\beta \|\mathrm{D}h(x)\|_{\mathrm{op}}^2}\right), \tag{3.5}$$

where $C_\lambda(x) = \|\mathrm{D}\lambda(x)\|_{\mathrm{op}}$ (Definition 2.1).

Proof Given $x \in \mathcal{C}$, consider the gradient step $x_t = x - t\nabla g(x)$ for some $t \geq 0$. We wish to find $t_{\max} > 0$ such that $x_t \in \mathcal{C}$ for all $t \in [0, t_{\max}]$. Using A3, we have

$$h(x_t) = h(x - t\nabla g(x)) \tag{3.6}$$

$$= h(x) - t\mathrm{D}h(x)[\nabla g(x)] + E(x, -t\nabla g(x)) \tag{3.7}$$

where $\|E(x, -t\nabla g(x))\| \leq C_h \|t\nabla g(x)\|^2$. Using Eq. (2.8) gives

$$h(x_t) = h(x) - t\mathrm{D}h(x)\big[\mathrm{grad}_{\mathcal{M}_x} f(x) + 2\beta \mathrm{D}h(x)^*[h(x)] - \mathrm{D}\lambda(x)^*[h(x)]\big] + E(x, -t\nabla g(x)). \tag{3.8}$$

Since $\mathrm{grad}_{\mathcal{M}_x} f(x)$ belongs to $\ker(\mathrm{D}h(x))$ by construction, one term cancels:

$$h(x_t) = h(x) - 2\beta t\mathrm{D}h(x)\big[\mathrm{D}h(x)^*[h(x)]\big] + t\mathrm{D}h(x)\big[\mathrm{D}\lambda(x)^*[h(x)]\big]$$

$$+ E(x, -t\nabla g(x)) \tag{3.9}$$
$$= \big(\mathrm{I}_m - 2\beta t \mathrm{D}h(x) \circ \mathrm{D}h(x)^*\big)[h(x)] + \big(t\mathrm{D}h(x) \circ \mathrm{D}\lambda(x)^*\big)[h(x)]$$
$$+ E(x, -t\nabla g(x)). \tag{3.10}$$

Let $\sigma_1 \geq \cdots > \sigma_m > 0$ denote the singular values of $\mathrm{D}h(x)$. The eigenvalues of the symmetric operator $(\mathrm{I}_m - 2\beta t \mathrm{D}h(x) \circ \mathrm{D}h(x)^*)$ are $1 - 2\beta t\sigma_1^2 \leq \cdots \leq 1 - 2\beta t\sigma_m^2$. All these eigenvalues are smaller than one and are nonnegative provided $0 \leq 1-2\beta t\sigma_1^2$ and $t \geq 0$, or equivalently:

$$0 \leq t \leq \frac{1}{2\beta \left\|\mathrm{D}h(x)\right\|_{\mathrm{op}}^2}. \tag{3.11}$$

Under that assumption, we further find:

$$\|h(x_t)\| \leq (1 - 2\beta t\sigma_m^2)\, \|h(x)\| + t\sigma_1 \left\|\mathrm{D}\lambda(x)^*\right\|_{\mathrm{op}} \|h(x)\| + C_h t^2 \left\|\nabla g(x)\right\|^2. \tag{3.12}$$

We want to show $\|h(x_t)\| \leq R$, which is indeed the case if

$$(1 - 2\beta t\sigma_m^2)\, \|h(x)\| + t\sigma_1 C_\lambda(x)\, \|h(x)\| + C_h \left\|\nabla g(x)\right\|^2 t^2 \leq R. \tag{3.13}$$

Thus, we seek conditions on t to ensure that the following quadratic inequality in t holds:

$$C_h \left\|\nabla g(x)\right\|^2 t^2 + \left(\sigma_1 C_\lambda(x) - 2\beta\sigma_m^2\right) \|h(x)\|\, t + \|h(x)\| - R \leq 0. \tag{3.14}$$

We branch into two cases. Firstly, consider $\|h(x)\| \in [R/2, R]$. In this case, (3.14) holds a fortiori if we remove the independent term $\|h(x)\| - R$ since the latter is nonpositive. By assumption, $\beta > \beta_1(x) = \sigma_1 C_\lambda(x)/2\sigma_m^2$, so the linear term is nonpositive. Therefore, we can upper bound the quadratic by setting $\|h(x)\| = R/2$. This shows that

$$C_h \left\|\nabla g(x)\right\|^2 t^2 + \left(\sigma_1 C_\lambda(x) - 2\beta\sigma_m^2\right) \|h(x)\|\, t + \|h(x)\| - R$$
$$\leq C_h \left\|\nabla g(x)\right\|^2 t^2 + \left(\sigma_1 C_\lambda(x) - 2\beta\sigma_m^2\right) \frac{R}{2} t. \tag{3.15}$$

The above is a convex quadratic with two real roots. It is nonpositive—and (3.14) is satisfied—if:

$$0 \leq t \leq \frac{\left(2\beta\sigma_m^2 - \sigma_1 C_\lambda(x)\right) R}{2C_h \left\|\nabla g(x)\right\|^2}. \tag{3.16}$$

For $\|h(x)\| \in [0, R/2]$, the linear term in (3.14) is still nonpositive. Additionally, the constant term of the quadratic is upper bounded by $-R/2$. This establishes

$$C_h \|\nabla g(x)\|^2 t^2 + \left(\sigma_1 C_\lambda(x) \|h(x)\| - 2\beta\sigma_m^2 \|h(x)\|\right) t + \|h(x)\| - R \\ \leq C_h \|\nabla g(x)\|^2 t^2 - \frac{R}{2}. \tag{3.17}$$

We infer that, for $\|h(x)\| \in [0, R/2]$, condition (3.14) is satisfied for

$$0 \leq t \leq \sqrt{\frac{R}{2C_h}} \frac{1}{\|\nabla g(x)\|}. \tag{3.18}$$

The main claim follows by collecting the conditions in Eqs. (3.11), (3.16) and (3.18). □

We now show that the backtracking in Algorithm 2 terminates in a finite number of steps and guarantees a sufficient decrease.

Lemma 3.2 *(Gradient step decrease) Take $x \in \mathcal{C}$ and $\beta > \beta_1(x)$. The backtracking procedure in Algorithm 2 terminates with a step-size $t \geq \tau_1 \min(\underline{\alpha_1}, t_1(x)) > 0$ where*

$$\underline{\alpha_1} = \min\left(\alpha_{01}, \frac{2(1-c_1)}{L_g}\right),$$

with $\alpha_{01} > 0$ the initial step size of Algorithm 2 and $t_1(x)$ is defined in Eq. (3.5). This guarantees the following decrease:

$$g(x) - g(x - t\nabla g(x)) \geq c_1\tau_1 \min(\underline{\alpha_1}, t_1(x)) \|\nabla g(x)\|^2. \tag{3.19}$$

Proof From Proposition 3.1, we know that $x - \alpha\nabla g(x)$ is in $\mathcal{C}$ for every $0 \leq \alpha \leq t_1(x)$. We proceed to show that the Armijo decrease condition (3.1) is satisfied for any $0 \leq \alpha \leq \min(t_1(x), \underline{\alpha_1})$. For every $0 \leq \alpha \leq t_1(x)$, the norm of the Hessian of g is bounded by the constant $\overline{L_g}$ (Eq. 3.3), which implies that ∇g is L_g-Lipschitz continuous on the segment that connects x and $x - t_1(x)\nabla g(x)$. Thus, for all $0 \leq \alpha \leq t_1(x)$, we have

$$g(x - \alpha\nabla g(x)) \leq g(x) + \langle -\alpha\nabla g(x), \nabla g(x)\rangle + \frac{L_g}{2} \|\alpha\nabla g(x)\|^2 \\ = g(x) + \left(\frac{\alpha L_g}{2} - 1\right)\alpha \|\nabla g(x)\|^2.$$

This is equivalent to $g(x) - g(x - \alpha\nabla g(x)) \geq \left(1 - \alpha L_g/2\right)\alpha \|\nabla g(x)\|^2$. For $0 \leq \alpha \leq 2(1-c_1)/L_g$, we have $(1 - \alpha L_g/2) \geq c_1$. Hence, for $0 \leq \alpha \leq \min(t_1(x), \underline{\alpha_1})$, condition (3.1), $g(x) - g(x - \alpha\nabla g(x)) \geq c_1\alpha \|\nabla g(x)\|^2$ is satisfied. Given $\beta > \beta_1(x)$, one readily checks that $t_1(x)$ is positive. Since $\underline{\alpha_1}$ is also positive, there exists a nonempty interval, $]0, \min(\underline{\alpha_1}, t_1(x))]$, where the step size satisfies the Armijo condition and defines a next iterate inside $\mathcal{C}$. Therefore, Algorithm 2 returns a step t satisfying $t \geq \tau_1 \min(\underline{\alpha_1}, t_1(x))$. In addition, the Armijo condition gives

$$\begin{aligned} g(x) - g(x - t\nabla g(x)) &\geq c_1 t \, \|\nabla g(x)\|^2 \\ &\geq c_1 \tau_1 \min(\underline{\alpha_1}, t_1(x)) \, \|\nabla g(x)\|^2 . \end{aligned}$$

□

3.3 Second-Order Steps

We begin with a result which guarantees small enough steps stay in $\mathcal{C}$ when $\nabla g(x)$ is small.

Proposition 3.3 *Suppose* A1 *and* A3 *hold. Take* $x \in \mathcal{C}$ *with* $\beta > \max\{\beta_1(x), \beta_2(x), \beta_3(x)\}$. *Assume that* $\|\nabla g(x)\| \leq \varepsilon_1$ *for some* $\varepsilon_1 \leq R/2$ *so that Proposition* 2.2 *applies. For any* $d \in \mathcal{E}$ *with* $\|d\| = 1$*, the point* $x + td$ *is in* $\mathcal{C}$ *for all* t *in the interval* $[0, t_2(x)]$ *with* $t_2(x)$ *defined by*

$$t_2(x) := \left(-\sigma_1(\mathrm{D}h(x)) + \sqrt{\sigma_1(\mathrm{D}h(x))^2 + 2C_h R}\right) / 2C_h. \tag{3.20}$$

Proof Since $\|\nabla g(x)\| \leq \varepsilon_1$, Proposition 2.2 ensures $\|h(x)\| \leq \varepsilon_1$. For $t > 0$, A3 yields

$$\begin{aligned} h(x + td) &= h(x) + t\mathrm{D}h(x)[d] + E(x, td), \\ \|h(x + td)\| &\leq \|h(x)\| + t\sigma_1 \, \|d\| + C_h t^2 \, \|d\|^2 \\ &\leq \varepsilon_1 + t\sigma_1 + C_h t^2, \end{aligned}$$

where σ_1 is the largest singular value of $\mathrm{D}h(x)$. We want to find the values of $t \geq 0$ for which $\varepsilon_1 + t\sigma_1 + C_h t^2 \leq R$. The convex quadratic $t \mapsto C_h t^2 + \sigma_1 t + \varepsilon_1 - R$ has roots $\left(-\sigma_1 \pm \sqrt{\sigma_1^2 - 4(\varepsilon_1 - R)C_h}\right) / 2C_h$, which for $\varepsilon_1 < R$ are real and of opposite signs. Hence, the quadratic is nonpositive for all t such that

$$0 \leq t \leq \left(-\sigma_1 + \sqrt{\sigma_1^2 + 4|R - \varepsilon_1|C_h}\right) / 2C_h.$$

By assumption, $\varepsilon_1 \leq R/2$ and therefore $x + td$ belongs to $\mathcal{C}$ for all t such that

$$0 \leq t \leq \left(-\sigma_1 + \sqrt{\sigma_1^2 + 4C_h R/2}\right) / 2C_h.$$

□

We now show that the backtracking of Algorithm 3 terminates in a finite number of steps and guarantees a sufficient decrease.

Lemma 3.4 *(*Eigenstep decrease*) Take $x \in \mathcal{C}$ and $\beta > \max\{\beta_1(x), \beta_2(x), \beta_3(x)\}$ with $\|\nabla g(x)\| \leq \varepsilon_1$ for some $\varepsilon_1 \leq R/2$. Assume there exists a direction $d \in \mathcal{E}$ such that $\|d\| = 1$, $\langle d, \nabla^2 g(x)[d]\rangle < -\varepsilon_2$ for some $\varepsilon_2 > 0$ and $\langle d, \nabla g(x)\rangle \leq 0$. The backtracking procedure in Algorithm 3 terminates with a step size $t \geq \tau_2 \min(\alpha_2(x), t_2(x)) > 0$ where*

$$\alpha_2(x) = \min\left(\alpha_{02}, \frac{3|2c_2 - 1||\langle d, \nabla^2 g(x)[d]\rangle|}{M_g}\right),$$

with $\alpha_{02} > 0$ the initial step size of Algorithm 3 and $t_2(x)$ is defined in Eq. (3.20). This ensures the following decrease:

$$g(x) - g(x + td) \geq -c_2\tau_2^2 \min(\alpha_2(x), t_2(x))^2 \langle d, \nabla^2 g(x)[d]\rangle. \tag{3.21}$$

Proof From Proposition 3.3, the point $x + \alpha d$ is in $\mathcal{C}$ for all $0 \leq \alpha \leq t_2(x)$. We show that for all $0 \leq \alpha \leq \min(\alpha_2(x), t_2(x))$, the decrease condition (3.2) is satisfied. For every $0 \leq \alpha \leq t_2(x)$, the norm of the third derivative of g is bounded by the constant M_g (Eq. 3.3), which implies that $\nabla^2 g$ is M_g-Lipschitz continuous on the segment that connects x and $x + t_2(x)d$. Thus, for all $0 \leq \alpha \leq t_2(x)$, we have

$$\begin{aligned} g(x + \alpha d) &\leq g(x) + \alpha\langle d, \nabla g(x)\rangle + \frac{\alpha^2}{2}\langle d, \nabla^2 g(x)[d]\rangle + \frac{M_g}{6}\alpha^3 \|d\|^3 \\ &\leq g(x) + \frac{\alpha^2}{2}\langle d, \nabla^2 g(x)[d]\rangle + \frac{M_g}{6}\alpha^3 \\ &\leq g(x) + \frac{\alpha^2}{2}\left(\langle d, \nabla^2 g(x)[d]\rangle + \frac{M_g\alpha}{3}\right). \end{aligned}$$

The sufficient decrease condition (3.2), $g(x) - g(x + \alpha d) \geq -c_2\alpha^2\langle d, \nabla^2 g(x)[d]\rangle$, is satisfied if

$$-\frac{\alpha^2}{2}\left(\langle d, \nabla^2 g(x)[d]\rangle + \frac{M_g\alpha}{3}\right) \geq -c_2\alpha^2\langle d, \nabla^2 g(x)[d]\rangle.$$

This is equivalent to

$$\begin{aligned} \langle d, \nabla^2 g(x)[d]\rangle + \frac{M_g\alpha}{3} &\leq 2c_2\langle d, \nabla^2 g(x)[d]\rangle \\ \alpha &\leq \frac{3(2c_2 - 1)\langle d, \nabla^2 g(x)[d]\rangle}{M_g} \\ \alpha &\leq \frac{3|2c_2 - 1||\langle d, \nabla^2 g(x)[d]\rangle|}{M_g}, \end{aligned}$$

since $c_2 < 1/2$. Therefore, (3.2) is satisfied for all $\alpha \leq \min(\alpha_2(x), t_2(x))$. One readily checks that $t_2(x)$ and $\alpha_2(x)$ are positive. Therefore, there exists a nonempty interval $(0, \min(\alpha_2(x), t_2(x))]$ where the step-size satisfies the decrease condition (3.2) and defines a next iterate inside $\mathcal{C}$. Therefore the backtracking in Algorithm 3 returns a step-size t satisfying $t \geq \tau_2 \min(\alpha_2(x), t_2(x))$ in a finite number of iterations. In addition, the decrease condition (3.2) provides

$$\begin{aligned} g(x) - g(x+td) &\geq -c_2 t^2 \langle d, \nabla^2 g(x)[d] \rangle \\ &\geq -c_2 \tau_2^2 \min(\alpha_2(x), t_2(x))^2 \langle d, \nabla^2 g(x)[d] \rangle. \end{aligned}$$

□

Remark 3.1 It may seem surprising that $\underline{\alpha_1}$ is a constant and $\alpha_2(x)$ depends on x through the quadratic term $|\langle d, \nabla^2 g(x)[d] \rangle|$. This is a consequence of the way first- and second-order directions are defined. The step-size for a first-order step multiplies the gradient which can vary in norm whereas the step-size in second-order steps always multiplies a unit-norm direction.

3.4 Worst-Case Global Complexity

We are now in a position to give a formal version of our main result: the worst-case complexity of the Gradient-Eigenstep algorithm for problem (P).

Theorem 3.5 *(Complexity of Algorithm 1) Consider Problem (P) under A1, A2, A3, A4 and A5. Let $0 < \varepsilon_1 \leq R/2$ and let $\underline{g}$ be the lower bound of g over the compact set $\mathcal{C}$. Algorithm 1 produces an iterate $x_{N_1} \in \mathcal{C}$ satisfying $\left\|\nabla g(x_{N_1})\right\| \leq \varepsilon_1$ with*

$$N_1 \leq \frac{g(x_0) - \underline{g}}{c_1 \tau_1 \min(\underline{\alpha_1}, \underline{t_1}) \varepsilon_1^2}, \tag{3.22}$$

where $\underline{t_1} = \min_{x \in \mathcal{C}} t_1(x) > 0$.
Furthermore if $0 < \varepsilon_2 < \infty$, Algorithm 1 also produces an iterate x_{N_2} satisfying $\left\|\nabla g(x_{N_2})\right\| \leq \varepsilon_1$ and $\lambda_{\min}\left(\nabla^2 g(x_{N_2})\right) \geq -\varepsilon_2$ with

$$N_2 \leq (g(x_0) - \underline{g}) \left[\min\left(c_1 \tau_1 \min(\underline{\alpha_1}, \underline{t_1}) \varepsilon_1^2, c_2 \tau_2^2 \min\left(\min\left(\alpha_{02}, \frac{3|2c_2 - 1|\varepsilon_2}{M_g} \right), \underline{t_2} \right)^2 \varepsilon_2 \right) \right]^{-1}, \tag{3.23}$$

where $\underline{t_2} = \min_{x \in \mathcal{C}} t_2(x) > 0$. The iterate x_{N_1} is an $(\varepsilon_1, 2\varepsilon_1)$-FOCP of (P) and x_{N_2} is an $(\varepsilon_1, 2\varepsilon_1, \varepsilon_2 + C\varepsilon_1)$-SOCP of (P), where C is defined in Corollary 2.7.

Proof We first show that the constants $\underline{t_1} = \min_{x \in \mathcal{C}} t_1(x)$ and $\underline{t_2} = \min_{x \in \mathcal{C}} t_2(x)$ are positive. Recall from Eq. (3.5) that

$$t_1(x) = \min\left(\sqrt{\frac{R}{2C_h}} \frac{1}{\|\nabla g(x)\|}, \frac{(2\beta \sigma_{\min}(\mathrm{D}h(x))^2 - \sigma_1(\mathrm{D}h(x)) C_\lambda(x)) R}{2C_h \|\nabla g(x)\|^2}, \right.$$

$$\left.\frac{1}{2\beta\,\|\mathrm{D}h(x)\|_{\mathrm{op}}^2}\right).$$

One readily checks that $t_1(x) > 0$ for all $x \in \mathcal{C}$. The first term, $\sqrt{R/2C_h}/\,\|\nabla g(x)\|$ is positive since ∇g is continuous over $\mathcal{C}$ and $\mathcal{C}$ is compact. Using that $\beta > \bar{\beta}_1$ (A5), the numerator of the second term is positive and bounded away from zero for all $x \in \mathcal{C}$. Using compactness of $\mathcal{C}$ and smoothness of h, the quantity $\|\mathrm{D}h(x)\|_{\mathrm{op}}$ is upper bounded over $\mathcal{C}$ and therefore $1/2\beta\,\|\mathrm{D}h(x)\|_{\mathrm{op}}^2$ is bounded away from zero over $\mathcal{C}$. We note that t_1 is a continuous function of x which is positive for all x in the compact set $\mathcal{C}$. Therefore, $\min_{x\in\mathcal{C}} t_1(x)$ is attained at a point in $\mathcal{C}$ and $\underline{t_1} > 0$. A similar process shows that $\underline{t_2} > 0$. The function $t_2(x) = \left(-\sigma_1(\mathrm{D}h(x)) + \sqrt{\sigma_1(\mathrm{D}h(x))^2 + 2C_h R}\right)/2C_h$ is continuous over $\mathcal{C}$. We also note that $t_2(x) > 0$ for all $x \in \mathcal{C}$ since the constants R and C_h are positive as a consequence of A1 and A3 respectively.

For every iterations k where a first-order step is performed, one has $\|\nabla g(x_k)\| > \varepsilon_1$, while for second-order steps $\langle d, \nabla^2 g(x_k)[d]\rangle < -\varepsilon_2$. Therefore, Eq. (3.19) gives the following decrease for first-order steps:

$$\begin{aligned} g(x_k) - g(x_{k+1}) &\geq c_1\tau_1 \min(\underline{\alpha_1}, t_1(x_k))\,\|\nabla g(x_k)\|^2 \\ &\geq c_1\tau_1 \min(\underline{\alpha_1}, \underline{t_1})\varepsilon_1^2, \end{aligned} \tag{3.24}$$

where $\underline{t_1} = \min_{x\in\mathcal{C}} t_1(x) > 0$, as shown above. The decrease for second-order steps follows from Eq. (3.21), that is,

$$\begin{aligned} g(x_k) - g(x_{k+1}) &\geq -c_2\tau_2^2 \min(\alpha_2(x_k), t_2(x_k))^2 \langle d, \nabla^2 g(x)[d]\rangle \\ &\geq c_2\tau_2^2 \min\left(\min\left(\alpha_{02}, \frac{3|2c_2-1||\langle d, \nabla^2 g(x_k)[d]\rangle|}{M_g}\right), t_2(x_k)\right)^2 \varepsilon_2 \\ &\geq c_2\tau_2^2 \min\left(\min\left(\alpha_{02}, \frac{3|2c_2-1|\varepsilon_2}{M_g}\right), \underline{t_2}\right)^2 \varepsilon_2, \end{aligned} \tag{3.25}$$

where $\underline{t_2} = \min_{x\in\mathcal{C}} t_2(x) > 0$, as shown above. Since $\mathcal{C}$ is compact (A2) and g is continuous on $\mathcal{C}$, let $\underline{g} := \min_{x\in\mathcal{C}} g(x) > -\infty$. Consider the case $\varepsilon_2 < \infty$. For any $K \geq 0$, we have

$$g(x_0) - \underline{g} \geq \sum_{k=0}^{K} g(x_k) - g(x_{k+1}) \tag{3.26}$$

$$\geq K \min\left(c_1\tau_1 \min(\underline{\alpha_1}, \underline{t_1})\varepsilon_1^2, c_2\tau_2^2 \min\left(\min\left(\alpha_{02}, \frac{3|2c_2-1|\varepsilon_2}{M_g}\right), \underline{t_2}\right)^2 \varepsilon_2\right). \tag{3.27}$$

Given the definition of N_2, Eq. (3.27) tells us that $K \leq N_2$. Hence, if more than N_2 iterations are performed, it must be that a point where $\|\nabla g(x)\| \leq \varepsilon_1$ and $\lambda_{\min}(\nabla^2 g(x)) \geq -\varepsilon_2$ has been encountered. In the case $\varepsilon_2 = \infty$, no second-order step is performed, which simplifies as follows:

$$\begin{aligned} g(x_0) - \underline{g} \geq \sum_{k=0}^{K} g(x_k) - g(x_{k+1}) \\ &\geq K c_1 \tau_1 \min(\underline{\alpha_1}, t_1(x_k)) \, \|\nabla g(x_k)\|^2 \\ &\geq K c_1 \tau_1 \min(\underline{\alpha_1}, \underline{t_1}) \varepsilon_1^2. \end{aligned} \tag{3.28}$$

The fact that x_{N_1} and x_{N_2} are respectively $(\varepsilon_1, 2\varepsilon_1)$-FOCP and $(\varepsilon_1, 2\varepsilon_1, \varepsilon_2 + C\varepsilon_1)$-SOCP of (P) follows from Propositions 2.2 and 2.6. □

4 Estimating the Penalty Parameter

The previous section establishes convergence results under the assumption that the penalty parameter β is large enough to satisfy A5. In practice, it is rarely possible to know whether this assumption is satisfied. Therefore, this section outlines a scheme which estimates a suitable value for β. A simple strategy consists in letting Algorithm 1 run for a fixed number of iterations using a given value of β, and to repeat this process with increasingly larger values of β until convergence of Algorithm 1 to an approximate critical point is achieved. We refer to such strategy as a plateau scheme. Along the way, if the algorithm encounters a point $x \in \mathcal{C}$ such that $\beta \leq \max\{\beta_1(x), \beta_2(x), \beta_3(x)\}$ (Definition 2.2), β is increased and the procedure is restarted. It is tedious but not difficult to show that the complexity of such a plateau scheme is only a logarithmic factor worse than the complexity of Algorithm 1—with a fixed value of β that satisfies A5.

Theorem 4.1 *Under A1, A2, A3 and A4, a plateau scheme returns an* $(\varepsilon_1, 2\varepsilon_1, \varepsilon_2 + C\varepsilon_1)$*-SOCP in at most* $\mathcal{O}\left(\max\left\{\varepsilon_1^{-2}, \varepsilon_2^{-3}\right\} \max\left\{\log_\gamma \varepsilon_1^{-2}, \log_\gamma \varepsilon_2^{-3}\right\}\right)$ *gradient steps and eigensteps on the function* g*, where* $\gamma > 1$ *is the growth rate of* β *between two plateaus and* C *is defined in Corollary 2.7.*

5 Conclusion and Discussion

In this work, we consider a penalty function (Fletcher's augmented Lagrangian) for optimization under smooth equality constraints. We establish connections between its approximate critical points and the approximate critical points of the original constrained problem (P). We also highlight that various definitions of approximate second-order critical points for equality constraints appear in the literature. Therefore, we propose a definition of approximate criticality which has a natural geometric interpretation and extends Riemannian optimality conditions to points near the feasible set.

We present Algorithm 1, which is shown to reach approximate second-order critical points of (P) in at most $\mathcal{O}(\varepsilon^{-3})$ iterations. The only other work to date which achieved this optimal rate for an infeasible method is [13], where the definition of approximate critical point is markedly different. Finally, we describe how Algorithm 1 can be modified to achieve a local quadratic convergence rate.

The main drawback of our approach, is the necessity to identify a set $\mathcal{C}$, where the differential of the constraint is nonsingular, in order to run the algorithm. Similar smoothness assumptions are made in related works which provide a worst-case complexity analysis [14, 40]. It would nonetheless be worthwhile to go beyond such assumptions.

Fletcher's augmented Lagrangian may be considered impractical in view of the linear system that must be solved at each iteration to evaluate the multipliers $\lambda(x)$. However, recent works show that it can still lead to the design of efficient algorithms and this work further reinforces the theoretical appeal of Fletcher's augmented Lagrangian. Directions of future research also emerge. Consider a smooth function $\hat{\lambda} \colon \mathcal{E} \to \mathbb{R}^m$ which coincides on $\mathcal{M}$ with the function $\lambda(x) = (\mathrm{D}h(x)^*)^\dagger [\nabla f(x)]$ considered in this work. This choice of multipliers defines a corresponding function $\hat{g}(x) = \mathcal{L}_\beta(x, \hat{\lambda}(x))$, a variant of the g. Recent works [21, 38, 39] show that minimizing the function $\hat{g}$ yields efficient algorithms for a particular choice of $\hat{\lambda}$ on the Stiefel manifold. Is there a way to generalize this concept to other manifolds? What theoretical guarantees can we hope to keep by using $\hat{\lambda}(x)$ instead of $\lambda(x)$? Exploring this could yield more practical Lagrangian-based infeasible methods to solve constrained optimization problems with underlying smoothness.

A Proof of Proposition 2.8

Proof Define $\varphi(x) = \frac{1}{2}\|h(x)\|^2$ and take any $x_0 \in \mathcal{C} = \{x \in \mathcal{E} : \varphi(x) \leq R^2/2\}$. Consider the following differential system:

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} x(t) = -\nabla\varphi(x(t)) \\ x(0) = x_0. \end{cases} \tag{A.1}$$

The fundamental theorem of flows [28, Theorem A.42] guarantees the existence of a unique maximal integral curve starting at x_0 for (A.1). Let $z(\cdot) \colon I \to \mathcal{E}$ denote this maximal integral curve and $T > 0$ be the supremum of the interval I on which $z(\cdot)$ is defined. We rely on the Escape Lemma [28, Lemma A.43] to show that $z(t)$ is defined for all times $t \geq 0$. For $t < T$, we write $\ell = \varphi \circ z$ and find

$$\ell'(t) = \mathrm{D}\varphi(z(t))\left[\frac{\mathrm{d}}{\mathrm{d}t} z(t)\right] = \left\langle \nabla\varphi(z(t)), \frac{\mathrm{d}}{\mathrm{d}t} z(t)\right\rangle \tag{A.2}$$

$$= -\|\nabla\varphi(z(t))\|^2 \tag{A.3}$$

$$= -\left\|\mathrm{D}h(z(t))^*[h(z(t))]\right\|^2 \leq 0. \tag{A.4}$$

This implies that $z(t) \in \mathcal{C}$ for all $0 \leq t < T$. We show that the trajectory $z(t)$ has finite length. To that end, we note that

$$\frac{1}{2}\|\nabla\varphi(x)\|^2 = \frac{1}{2}\left\|\mathrm{D}h(x)^*[h(x)]\right\|^2 \geq \underline{\sigma}^2\frac{1}{2}\|h(x)\|^2 = \underline{\sigma}^2\varphi(x), \tag{A.5}$$

for all $x \in \mathcal{C}$. The length of the trajectory from time $t = 0$ to $t = T$ is bounded as follows, using a classical argument [30]:

$$\begin{aligned}
\int_0^T \left\|\frac{\mathrm{d}}{\mathrm{d}t}z(t)\right\| \mathrm{d}t &= \int_0^T \|-\nabla\varphi(z(t))\| \,\mathrm{d}t \\
&= \int_0^T \frac{\|\nabla\varphi(z(t))\|^2}{\|\nabla\varphi(z(t))\|}\mathrm{d}t \\
&= \int_0^T \frac{\left\langle -\nabla\varphi(z(t)), \frac{\mathrm{d}}{\mathrm{d}t}z(t)\right\rangle}{\|\nabla\varphi(z(t))\|}\mathrm{d}t \\
&= \int_0^T \frac{-(\varphi\circ z)'(t)}{\|\nabla\varphi(z(t))\|}\mathrm{d}t \\
&\leq \int_0^T \frac{-(\varphi\circ z)'(t)}{\underline{\sigma}\sqrt{2(\varphi\circ z)(t)}}\mathrm{d}t \\
&= \frac{-\sqrt{2}}{\underline{\sigma}}\left[\sqrt{\varphi(z(T))} - \sqrt{\varphi(z(0))}\right] \\
&\leq \frac{\sqrt{2\varphi(z(0))}}{\underline{\sigma}}.
\end{aligned} \tag{A.6}$$

The length is bounded independently of T and therefore the flow has finite length. The Escape Lemma states that for a maximum integral curve $z(\cdot)\colon I \to \mathcal{E}$, if I has a finite upper bound, then the curve $z(\cdot)$ must be unbounded. Since $z(\cdot)$ is contained in a compact set by (A.6), the converse ensures that the interval I does not have a finite upper bound and therefore, $I = \mathbb{R}_+$. Since the trajectory $z(t)$ is bounded for $t \geq 0$, it must have an accumulation point $\bar{z}$. From A1, we have $\sigma_{\min}(\mathrm{D}h(z(t)) \geq \underline{\sigma} > 0$ for all $t \geq 0$. This gives the bound $\ell'(t) \leq -\underline{\sigma}^2\|h(z(t))\|^2 = -2\underline{\sigma}^2\ell(t)$. Gronwall's inequality then yields

$$\ell(t) \leq \varphi(x_0)e^{-2\underline{\sigma}^2 t}. \tag{A.7}$$

Therefore $\ell(t) \to 0$ as $t \to \infty$, which implies $h(z(t)) \to 0$ as $t \to \infty$. We conclude that the accumulation point satisfies $h(\bar{z}) = 0$. Since $\mathcal{C}$ is closed, the point $\bar{z}$ is in $\mathcal{C}$. Therefore, $\bar{z}$ is both in $\mathcal{M}$ and in the connected component of $\mathcal{C}$ that contains $z(0) = x_0$.
□

References

1. Ablin, P., Peyré, G.: Fast and accurate optimization on the orthogonal manifold without retraction. In International Conference on Artificial Intelligence and Statistics, pp. 5636–5657. PMLR (2022)
2. Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2008). https://doi.org/10.1515/9781400830244
3. Andreani, R., Martínez, J.M., Schuverdt, M.L.: On second-order optimality conditions for nonlinear programming. Optimization **56**(5–6), 529–542 (2007). https://doi.org/10.1080/02331930701618617
4. Bai, Y., Mei, S.: Analysis of Sequential Quadratic Programming Through the Lens of Riemannian Optimization. arXiv preprint arXiv:1805.08756 (2018)
5. Bai, Y., Duchi, J., Mei, S.: Proximal Algorithms for Constrained Composite Optimization, with Applications to Solving Low-Rank SDPs. arXiv preprint arXiv:1903.00184 (2019)
6. Bento, G.C., Ferreira, O.P., Melo, J.G.: iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. J. Optim. Theory Appl. **173**(2), 548–562 (2017). https://doi.org/10.1007/s10957-017-1093-4
7. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Academic Press, Cambridge (1982). https://doi.org/10.1016/C2013-0-10366-2
8. Birgin, E.G., Martínez, J.M.: Complexity and performance of an augmented Lagrangian algorithm. Optim. Methods Softw. **35**(5), 885–920 (2020). https://doi.org/10.1080/10556788.2020.1746962
9. Boumal, N., Absil, P.-A., Cartis, C.: Global rates of convergence for nonconvex optimization on manifolds. IMA J. Numer. Anal. **39**(1), 1–33 (2019). https://doi.org/10.1093/imanum/drx080
10. Boumal, Nicolas: An Introduction to Optimization on Smooth Manifolds. Cambridge University Press, Cambridge (2023). https://doi.org/10.1017/9781009166164
11. Burer, S., Monteiro, R.D.C.: A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. Math. Program. **95**(2), 329–357 (2003). https://doi.org/10.1007/s10107-002-0352-8
12. Cartis, C., Gould, N.I.M., Toint, Ph.L.: Complexity bounds for second-order optimality in unconstrained optimization. J. Complex. **28**(1), 93–108 (2012). https://doi.org/10.1016/j.jco.2011.06.001
13. Cartis, C., Gould, N.I.M., Toint, Ph.L.: Optimality of orders one to three and beyond: characterization and evaluation complexity in constrained nonconvex optimization. J. Complex. **53**, 68–94 (2019). https://doi.org/10.1016/j.jco.2018.11.001
14. Cifuentes, D., Moitra, A.: Polynomial time guarantees for the Burer–Monteiro method. Adv. Neural Inf. Process. Syst. **35**, 23923–23935 (2022)
15. Di Pillo, G.: Exact penalty methods. In: Algorithms for Continuous Optimization, pp. 209–253. Springer, Dordrecht (1994). https://doi.org/10.1007/978-94-009-0369-2_8
16. Di Pillo, G., Grippo, L.: An exact penalty function method with global convergence properties for nonlinear programming problems. Math. Program. **36**(1), 1–18 (1986). https://doi.org/10.1007/BF02591986
17. Di Pillo, G., Grippo, L.: Exact penalty functions in constrained optimization. SIAM J. Control. Optim. **27**(6), 1333–1360 (1989). https://doi.org/10.1137/0327068
18. Estrin, R., Friedlander, M.P., Orban, D., Saunders, M.A.: Implementing a smooth exact penalty function for equality-constrained nonlinear optimization. SIAM J. Sci. Comput. **42**(3), A1809–A1835 (2020). https://doi.org/10.1137/19M1238265
19. Estrin, R., Friedlander, M.P., Orban, D., Saunders, M.A.: Implementing a smooth exact penalty function for general constrained nonlinear optimization. SIAM J. Sci. Comput. **42**(3), A1836–A1859 (2020). https://doi.org/10.1137/19M1255069
20. Fletcher, R.: A class of methods for nonlinear programming with termination and convergence properties. In: Integer and nonlinear programming, pp. 157–173. Amsterdam (1970)
21. Gao, B., Liu, X., Yuan, Y.-X.: Parallelizable algorithms for optimization problems with orthogonality constraints. SIAM J. Sci. Comput. **41**(3), A1949–A1983 (2019). https://doi.org/10.1137/18M1221679
22. Ge, R., Huang, F., Jin, C., Yuan, Y.: Escaping from saddle points—online stochastic gradient for tensor decomposition. In: Proceedings of The 28th Conference on Learning Theory, pp. 797–842. PMLR (2015)
23. Goyens, F., Eftekhari, A., Boumal, N.: Computing second-order points under equality constraints: revisiting Fletcher's augmented Lagrangian. arXiv preprint arXiv:2204.01448 (2022)
24. Grapiglia, G.N., Yuan, Y.-X.: On the complexity of an augmented Lagrangian method for nonconvex optimization. IMA J. Numer. Anal. **41**(2), 1508–1530 (2021). https://doi.org/10.1093/imanum/draa021

25. Grubišić, I., Pietersz, R.: Efficient rank reduction of correlation matrices. Linear Algebra Appl. **422**(2), 629–653 (2007). https://doi.org/10.1016/j.laa.2006.11.024
26. He, C., Lu, Z., Pong, T. K.: A Newton-CG based augmented Lagrangian method for finding a second-order stationary point of nonconvex equality constrained optimization with complexity guarantees. arXiv preprint arXiv:2301.03139 (2023)
27. Horn, R.A., Johnson, C.R.: Topics in Matrix Analysis. Cambridge University Press, Cambridge (1991). https://doi.org/10.1017/CBO9780511840371
28. Lee, John M.: Introduction to Riemannian Manifolds, vol. 2. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91755-9
29. Ling, S.: Solving orthogonal group synchronization via convex and low-rank optimization: tightness and landscape analysis. Math. Program. **200**(1), 589–628 (2023). https://doi.org/10.1007/s10107-022-01896-3
30. Łojasiewicz, S.: Sur les trajectoires du gradient d'une fonction analytique. Seminari di geometria, pp. 115–117, (1982)
31. Nesterov, Yurii: Introductory Lectures on Convex Optimization. Springer, New York (2004). https://doi.org/10.1007/978-1-4419-8853-9
32. Polyak, B.T.: Gradient methods for minimizing functionals. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki **3**(4), 643–653 (1963). https://doi.org/10.1016/0041-5553(63)90382-3
33. Polyak, R.A.: On the local quadratic convergence of the primal-dual augmented Lagrangian method. Optim. Methods Softw. **24**(3), 369–379 (2009). https://doi.org/10.1080/10556780802699433
34. Rosen, D.M., Doherty, K.J., Terán Espinoza, A., Leonard, J.J.: Advances in inference and representation for simultaneous localization and mapping. Annu. Rev. Control Robot. Auton. Syst. **4**(1), 215–242 (2021). https://doi.org/10.1146/annurev-control-072720-082553
35. Royer, C.W., O'Neill, M., Wright, S.J.: A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. Math. Program. **180**(1), 451–488 (2020). https://doi.org/10.1007/s10107-019-01362-7
36. Schechtman, S., Tiapkin, D., Muehlebach, M., Moulines, E.: Orthogonal Directions Constrained Gradient Method: From non-linear equality constraints to Stiefel manifold. arXiv preprint arXiv:2303.09261 (2023)
37. Wright, S.J., Recht, B.: Optimization for Data Analysis. Cambridge University Press, Cambridge (2022). https://doi.org/10.1017/9781009004282
38. Xiao, N., Liu, X.: Solving optimization problems over the Stiefel manifold by smooth exact penalty function. arXiv preprint arXiv:2110.08986 (2021)
39. Xiao, N., Liu, X., Yuan, Y.-X.: A class of smooth exact penalty function methods for optimization problems with orthogonality constraints. Optim. Methods Softw. **37**(4), 1205–1241 (2022). https://doi.org/10.1080/10556788.2020.1852236
40. Xie, Y., Wright, S.J.: Complexity of proximal augmented Lagrangian for nonconvex optimization with nonlinear equality constraints. J. Sci. Comput. **86**(3), 1–30 (2021). https://doi.org/10.1007/s10915-021-01409-y
41. Zhang, H., Sra, S.: First-order methods for geodesically convex optimization. In: Conference on Learning Theory, pp. 1617–1638. PMLR (2016)