Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary

G. Haeser* Hongcheng Liu† Yinyu Ye‡

February 14, 2017

Abstract

In this paper we consider the minimization of a continuous function that is potentially not differentiable or not twice differentiable on the boundary of the feasible region. By exploiting an interior point technique, we present first- and second-order optimality conditions for this problem that reduces to classical ones when the derivative on the boundary is available. For this type of problems, existing necessary conditions often rely on the notion of subdifferential or become non-trivially weaker than the KKT condition in the (twice-)differentiable counterpart problems. In contrast, this paper presents a new set of first- and second-order necessary conditions that are derived without the use of subdifferential and reduces to exactly the KKT condition when (twice-)differentiability holds. As a result, these conditions are stronger than some existing ones considered for the discussed minimization problem when only non-negativity constraints are present. To solve for these optimality conditions in the special but important case of linearly constrained problems, we present two novel interior trust-region point algorithms and show that their worst-case computational efficiency in achieving the potentially stronger optimality conditions match the best known complexity bounds. Since this work considers a more general problem than the literature, our results also indicate that best known complexity bounds hold for a wider class of nonlinear programming problems.

Keywords: Constrained optimization, Nonconvex programming, Interior point method, First order algorithm, Nonsmooth problems

*Department of Applied Mathematics, University of São Paulo, São Paulo SP, Brazil. Visiting Scholar at Department of Management Science and Engineering, Stanford University, Stanford CA 94305, USA. E-mail: ghaeser@ime.usp.br.

†Department of Radiation Oncology, Stanford University, Stanford CA 94305, USA. E-mail: hql5143liu@gmail.com

‡Department of Management Science and Engineering, Stanford University, Stanford CA 94305, USA. E-mail: yinyu-ye@stanford.edu

1 Introduction

In this paper we are interested in the problem

$$\begin{array}{ll} \text{Minimize} & f(x), \\ \text{subject to} & \mathbf{A}x = \mathbf{b}, x \geq 0, \end{array} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $f : \mathbb{R}^n_+ \to \mathbb{R}$ is a continuous function on $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n \mid x \geq 0\}$ and smooth on $\mathbb{R}^n_{++} := \{x \in \mathbb{R}^n \mid x > 0\}$. As a special case of (1), the following formulation has been popularly studied:

$$\begin{array}{ll} \text{Minimize} & H(x) + \lambda \sum_{i=1}^n \varphi(x_i^p), \\ \text{subject to} & x \geq 0, \end{array} \tag{2}$$

where H is smooth, φ is convex, $\lambda > 0$ and $0 < p < 1$. A common use of (2) (or its immediate reformulations) is the problem of high-dimensional learning under the assumption of sparsity. In such a problem, few data observations are acquired for the task of recovering a high-dimension signal. Such a task is often done by minimizing an in-sample statistical loss (a.k.a., fidelity) function $H(x)$ that represents the in-sample error plus a regularization function $\lambda \sum_{i=1}^n \varphi(x_i^p)$, which penalizes non-zero variables to induce sparsity. Theoretical and numerical studies on the efficacies of this type of models are presented in [45, 41, 28, 29, 30, 31, 43, 52, 53]. Particularly, it is shown by [41, 43, 53, 31, 29, 52] that to achieve a sound recovery quality, global optimality to (1) is not necessary, but some local minima or even stationary points can successfully recover the high-dimensional signal with high probability. In specific, [41] shows that solutions satisfying a second-order necessary condition in linear regression penalized by certain nonconvex $\varphi(x_i^p)$ have very desirable statistical properties. [38] presented a recent application of (2) in designing neural networks for deep learning, for which $\varphi(x_i^p) = |x|$ or $\varphi(x_i^p) = \|x\|^2$ and H is a nonconvex loss function.

Despite various successful and seminal applications, (2) remains a non-trivial problem to solve due to the usual absence of differentiability or twice-differentiability and the frequent presence of nonconvexity. As an example, if $p < 1$, the function $\sum_{i=1}^n x_i^p$ is not even directionally differentiable in Gâteaux sense when $x_i = 0$ for any i. Similarly, when $p < 2$, the objective function is not twice differentiable. Meanwhile, in the training of a neural network, H is usually smooth but nonconvex, as in the case of [38]. [53] discussed some other cases where H is nonconvex.

To establish first-/second-order necessary optimality conditions for local minimality, different variants of the KKT condition have been discussed when differentiability is potentially absent. In such a case, optimality conditions based on the notion of subdifferential are studied by [25, 48, 1, 39]. Weaker optimality conditions without the use of subdifferential have been discussed by [11, 13, 12, 42]. Interested readers are referred to [10] for an excellent review on the optimality conditions. In particular, [13] considers the so-called scaled

first-order optimality condition for (2):

$$x_i \frac{\partial H(x)}{\partial x_i} + \lambda p \varphi'(x_i^p) x_i^p = 0, \qquad \forall i = 1, \ldots, n. \tag{3}$$

This condition is evidently weaker than the conditions by [25, 48, 1, 39], in that (3) always holds at the origin regardless of the objective function. According to [10], similar issues apply to the optimality conditions in [11, 12, 42]. In contrast, our presented optimality condition does not rely on any form of subdifferential and is equivalent to the canonical version of the KKT condition when f is smooth. Therefore, the presented optimality condition is tighter than [11, 13, 12, 42].

Our research is also motivated by the need of characterizing approximations to the "exact" necessary condition, since it is generally impossible to solve (1) exactly, even only for KKT solutions. As a result, the "exact" first- or second-order necessary conditions must be perturbed to properly characterize the actual solution obtained through an algorithm. Furthermore, it is desirable to establish a connection between the optimality condition and its ε perturbed version (approximation with inaccuracy measured by ε) in order for the complexity results to be meaningful. Approximate KKT-like conditions in solving nonconvex and nonsmooth optimization have been proposed by [13, 12, 25, 10]. In view of this gap in the literature, this paper presents a set of perturbed (first- and second-order) necessary optimality conditions that are originally defined in terms of a limit of perturbed stationary points. Compare to [25, 10], our perturbed necessary conditions are free from the use of subdifferential, and are stronger than [13, 12].

To compute solutions satisfying our proposed perturbed necessary conditions, we develop a first- and second-order interior trust-region point (ITRP) algorithms. Both algorithms work in a general setting that allows for irregularities of the objective function unaddressed in the literature. In particular, the first-order ITRP allows f to be not even directionally differentiable. The resulting computational complexity, $O(\varepsilon^{-2})$ in achieving an ε-perturbed first-order stationary point (where $\varepsilon > 0$), coincides with the best known complexity for solving smooth nonconvex problems using only first-order information and assuming the absence of matrix inversion. The second-order ITRP then applies to a class of problems where second-order derivative may not exist. The resulting complexity, $O(\varepsilon^{-3/2})$ and $O(\varepsilon^{-3})$ in achieving an ε-perturbed first-order and second-order stationary point, respectively, equals the best-known complexity for twice continuously differentiable functions. The corresponding ε-perturbed necessary optimality conditions are in stronger forms than those discussed in [13, 12, 25, 10]. We further show that, at the same rate of complexity, the same type of ε-perturbed scaled optimality condition as in [13] can be achieved for a more general set of optimization problems by our second-order ITRP. For a comprehensive analysis of the ITRP, we further considered the case where f is a quadratic function and present an alternative analysis for the same result in [55]. In such a special case, the ITRP is substantially accelerated and achieves both the first- and second-order conditions at a rate of $O(\varepsilon^{-1})$.

In contrast, in the literature, for smooth unconstrained optimization, when only first-order information is accessible and no matrix inversion is involved, the algorithms with best known complexity bounds take at most $O(\varepsilon^{-2})$ iterations to achieve a first-order stationary point up to a tolerance ε. It is the case of the steepest descent [46], trust region methods [36] and the nonlinear stepsize control algorithms [35, 49], for instance. When second-order derivatives are used, the best known complexity is reduced to $O(\varepsilon^{-3/2})$ for first-order stationarity and, to find a second-order stationary point perturbed by ε, the best known complexity is $O(\varepsilon^{-3})$. See [35, 49, 16, 27, 18, 24, 47, 44]. A different line of reasoning appeared recently in [17, 2], where the second-order information is iteratively approximated by the first-order one. In this case, the complexity bound of $O(\varepsilon^{-7/4})$ can be achieved for first-order stationarity. We do not pursue this last type of results. The best complexity bounds known are the same if constraints are considered [20, 21] or in some nonsmooth cases [11, 13, 12, 19, 34]. Our algorithms will achieve the best known complexity bounds of $O(\varepsilon^{-2})$, $O(\varepsilon^{-3/2})$ and $O(\varepsilon^{-3})$, depending on the use of second-order information. To our knowledge, our problem of discussion is more general than most existing developments in the literature.

The rest of the paper is organized in the following way. Section 2 articulates our optimality condition and Section 3 presents our algorithm and complexity analyses. Finally, Section 4 concludes the paper.

Notation. Given $n \geq 1$, $\mathbb{R}^n_+$ is the non-negative orthant in $\mathbb{R}^n$. We denote by $\mathbb{R}^n_{++} \subset \mathbb{R}^n_+$ the subset of vectors with all coordinates positive. Given $x \in \mathbb{R}^n$, we denote $diag(x)$ the diagonal matrix defined by x. When it is clear from confusion, we call $X = diag(x)$. The vectors $e_1, \ldots, e_n$ is the canonical basis of $\mathbb{R}^n$ and $e \in \mathbb{R}^n$ is the vector of ones. The identity matrix of appropriate dimention will be denoted $\mathcal{I}$. Given a symmetric matrix A, we denote by $A \succeq 0$ when A is positive semidefinite. The gradient vector and hessian matrix of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ is denoted, respectively, by $\nabla f(x)$ and $\nabla^2 f(x)$. We use $\|\cdot\|$ and $\|\cdot\|_\infty$ to represent the ℓ_2- and ℓ_∞-norms, respectively. The smallest integer greater than or equal to $x \in \mathbb{R}$ is denoted by $\lceil x \rceil$.

2 Optimality condition

Let us consider, for simplicity, a special case of (1) with only bound constraints $x \geq 0$ and let us assume that for each $i = 1, \ldots, n$, the partial derivative $\frac{\partial f(x)}{\partial x_i}$ is not defined when $x_i = 0$. A so-called scaled first-order optimality condition holds at a local minimizer x^*, given by $x_i^* \frac{\partial f(x^*)}{\partial x_i} = 0, i = 1, \ldots, n$, where the product is taken to be zero when the derivative does not exist. See [26].

A point $x > 0$ with $|x_i \frac{\partial f(x)}{\partial x_i}| \leq \varepsilon$ for all $i = 1, \ldots, n$, is called an ε-scaled first-order point. See [13]. In [12], it was proved that if a sequence $\{x^k\} \subset \mathbb{R}^n$ is such that $x^k \to x^*$ and x^k is an ε_k-scaled first-order point for all k with some $\varepsilon_k \to 0^+$, then x^* is a scaled first-order point. Combining both results,

the situation is the one described in Figure 1. Algorithms thus proceed to find ε-scaled first-order points, with some small $\varepsilon > 0$ as in [13, 12, 42].

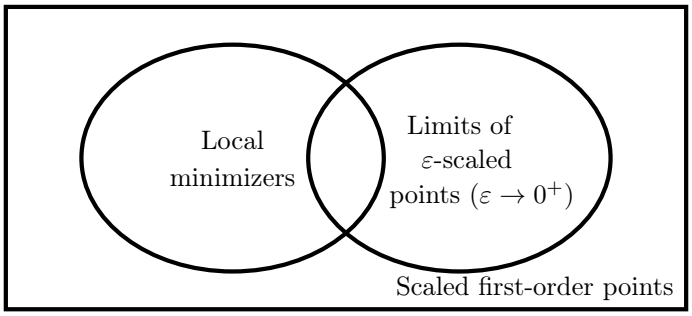

Figure 1: Local minimizers and limits of ε-scaled first-order points, $\varepsilon \to 0^+$, are scaled first-order points. Since a scaled first-order point can be seen as a weak necessary optimality condition, this gives little theoretical justification for considering an ε-scaled first-order point, $\varepsilon > 0$, as an approximate solution.

A first issue with this approach is that there is no analogous of the condition $\nabla f(x) \geq 0$, present in the canonical KKT conditions when derivatives exist everywhere. This is overcome in [13, 12, 42] by considering the particular objective function (2), where $\frac{\partial f(x)}{\partial x_i} \to +\infty$ when $x_i \to 0$, or considering an optimality condition based on the computation of subdifferentials [10]. A second issue is the fact that there is no measure of strength of the scaled first-order optimality condition, since, for instance, it always holds at $x = 0$, regardless of the objective function. Finally, a third issue is the lack of relation between local minimizers and limits of ε-scaled first-order points, as suggested by Figure 1. A similar criticism apply to the scaled second-order condition considered in [13], and other first-order optimality conditions considered for this class of problems. See [10] and references therein.

We will overcome these issues by defining first- and second-order optimality conditions that coincide with the canonical first- and second-order KKT conditions under usual smoothness assumptions, in a much more general framework. The optimality condition is defined in such a way that it naturally suggest an ε perturbed first- and second-order criterion suitable for the complexity analysis. We also show that, in the case of linear constraints, our first-order (second-order) optimality condition can be satisfied by the computation of ε-scaled first-order (second-order, respectively) points, as long as a suitable non-negativity criterion associated with the gradient of the objective function is fulfilled.

2.1 Necessary Optimality Conditions Based on Limits of Perturbations

This section presents optimality conditions for a much more general problem than (1). Specifically, we consider the problem:

$$\begin{array}{ll} \text{Minimize} & f(x), \\ \text{subject to} & h(x) = 0, c(x) \geq 0, \end{array} \tag{4}$$

where, $f : \mathbb{R}^n \to \mathbb{R}, h : \mathbb{R}^n \to \mathbb{R}^m$ and $c : \mathbb{R}^n \to \mathbb{R}^p$. Defining $C^\circ := \{x \mid c(x) > 0\}$ and $C := \{x \mid c(x) \geq 0\}$, f, h and c are assumed to be continuous on C and differentiable on C°. For the second-order optimality condition, we assume also second-order differentiability on C°. For any local solution x^* of (4), assume that there exists a sequence $\{z^k\}$ with $z^k \to x^*$ and $z^k \in C^\circ \cap \{x \mid h(x) = 0\}$ for all k, which is typically necessary for the application of interior point methods. Also assume that for any point $x \in C^\circ \cap \{x \mid h(x) = 0\}$, the rank of $\{\nabla h_i(y)\}_{i=1}^m$ is constant for all y in a neighborhood of x.

Note that derivatives of objective function and constraints may not exist when some $c_i(x) = 0$. Note also that we do not assume any constraint qualification on the whole feasible set.

Theorem 1. *Under the assumptions described above, let x^* be a local solution of (4). Then, there exists a sequence of approximate solutions $\{x^k\} \subset \mathbb{R}^n$ and sequences of approximate Lagrange multipliers $\{\lambda^k\} \subset \mathbb{R}^m$, $\{s^k\} \subset \mathbb{R}^p_+$ such that:*

i) $c(x^k) > 0$, $h(x^k) = 0$ *for all k and* $x^k \to x^*$,

ii) $\nabla f(x^k) + \sum_{i=1}^m \lambda_i^k \nabla h_i(x^k) - \sum_{i=1}^p s_i^k \nabla c_i(x^k) \to 0$,

iii) $c_i(x^k) s_i^k \to 0$ *for all* $i = 1, \ldots, p$.

If, in addition, f, h, and c are twice differentiable on C°, then, there exist sequences $\{\theta^k\} \subset \mathbb{R}^p_+$ and $\{\delta_k\} \subset \mathbb{R}_+, \delta_k \to 0^+$ such that

iv) $d^\top(\nabla^2 f(x^k) + \sum_{i=1}^m \lambda_i^k \nabla^2 h_i(x^k) - \sum_{i=1}^p s_i^k \nabla^2 c_i(x^k) + \sum_{i=1}^p \theta_i^k \nabla c_i(x^k) \nabla c_i(x^k)^\top + \delta_k \mathcal{I}) d \geq 0$, *for all $d \in \mathbb{R}^n$ with $\nabla h_i(x^k)^\top d = 0, i = 1, \ldots, m$.*

v) $c_i(x^k)^2 \theta_i^k \to 0^+$ *for all* $i = 1, \ldots, p$.

Proof. Let us take $\delta > 0$ small enough such that the problem

$$\text{Minimize } f(x) + \frac{1}{4}\|x - x^*\|^4, \text{ s.t. } c(x) \geq 0, h(x) = 0, \|x - x^*\|^2 \leq \delta, \tag{5}$$

has x^* as its unique global solution.

Let us consider the application of the classical interior penalty method [32] to problem (5) in the following sense: given a sequence $\{\mu_k\} \subset \mathbb{R}_+, \mu_k > 0$ with $\mu_k \to 0^+$, consider for every k the problem:

$$\begin{array}{ll} \text{Minimize} & \varphi_k(x) := f(x) + \frac{1}{4}\|x - x^*\|^4 - \mu_k \sum_{i=1}^m \log(c_i(x)), \\ \text{subject to} & c(x) > 0, h(x) = 0, \|x - x^*\|^2 \leq \delta. \end{array} \tag{6}$$

It is well known that a global solution x^k exists for all k and that cluster points of $\{x^k\}$ are global solutions of (5), see [32]. By the last constraint of (6), $\{x^k\}$ is bounded, which implies that $x^k \to x^*$ and thus i) holds.

For k large enough, x^k is a local solution of

$$\text{Minimize } \varphi_k(x) := f(x) + \frac{1}{4}\|x - x^*\|^4 - \mu_k \sum_{i=1}^m \log(c_i(x)), \text{ s.t. } h(x) = 0.$$

Since the constraints $h(x) = 0$ satisfy a constraint qualification, there exist Lagrange multipliers $\lambda^k \in \mathbb{R}^m$ such that

$$\begin{aligned} 0 =& \nabla \varphi_k(x^k) + \sum_{i=1}^m \lambda_i^k \nabla h_i(x^k) \\ =& \nabla f(x^k) + \|x^k - x^*\|^2(x^k - x^*) + \sum_{i=1}^m \lambda_i^k \nabla h_i(x^k) - \sum_{i=1}^p \frac{\mu_k}{c_i(x^k)} \nabla c_i(x^k), \end{aligned}$$

which gives ii) and iii) for $s_i^k := \frac{\mu_k}{c_i(x^k)}, i = 1, \ldots, p$.

The second-order differentiability assumption and the constant rank condition around x^k is enough to ensure that (see [9]):

$$\begin{aligned} 0 \leq & d^\top (\nabla^2 \varphi(x^k) + \sum_{i=1}^m \lambda_i^k \nabla^2 h_i(x^k)) d \\ = & d^\top \left(\nabla^2 f(x^k) + \sum_{i=1}^m \lambda_i^k \nabla^2 h_i(x^k) - \sum_{i=1}^p s_i^k \nabla^2 c_i(x^k) \right. \\ & \left. + \sum_{i=1}^p \frac{\mu_k}{c_i(x^k)^2} \nabla c_i(x^k) \nabla c_i(x^k)^\top + 2(x^k - x^*)(x^k - x^*)^\top + \|x^k - x^*\|^2 \mathcal{I} \right) d, \end{aligned}$$

for all $d \in \mathbb{R}^n$ such that $\nabla h_i(x^k)^\top d = 0, i = 1, \ldots, m$.

The result follows defining $\theta_i^k := \frac{\mu_k}{c_i(x^k)^2}$ for all $i = 1, \ldots, p$, and $\delta^k \geq 0$ as the largest eigenvalue of $2(x^k - x^*)(x^k - x^*)^\top + \|x^k - x^*\|^2 \mathcal{I}$ for all k, which converges to zero. $\square$

The optimality conditions immediately suggests definitions for ε-perturbed first- and second-order stationary points:

Definition 1. Given $\varepsilon > 0$, a point $x \in \mathbb{R}^n$ is called an ε-KKT point for problem (4) when there exist approximate Lagrange multipliers $\lambda \in \mathbb{R}^m$ and $s \in \mathbb{R}^p_+$ with:

(i) $h(x) = 0$, $c(x) > 0$,

(ii) $\|\nabla f(x) + \sum_{i=1}^m \lambda_i \nabla h_i(x) - \sum_{i=1}^p s_i \nabla c_i(x)\|_\infty \leq \varepsilon$,

(iii) $|c_i(x) s_i| \leq \varepsilon$ for all $i = 1, \ldots, p$.

Definition 2. Given $\varepsilon > 0$, a point $x \in \mathbb{R}^n$ is called an ε-KKT2 point for problem (4) when there exist approximate Lagrange multipliers $\lambda \in \mathbb{R}^m$ and $s \in \mathbb{R}^p_+$ and a parameter $\theta \in \mathbb{R}^p_+$ with:

(i) $h(x) = 0$, $c(x) > 0$,

(ii) $\|\nabla f(x) + \sum_{i=1}^m \lambda_i \nabla h_i(x) - \sum_{i=1}^p s_i \nabla c_i(x)\|_\infty \leq \varepsilon$,

(iii) $|c_i(x) s_i| \leq \varepsilon$ for all $i = 1, \ldots, p$,

(iv) $d^\top \left(\nabla^2 f(x) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x) - \sum_{i=1}^p s_i \nabla^2 c_i(x) + \sum_{i=1}^p \theta_i \nabla c_i(x) \nabla c_i(x)^\top + \varepsilon \mathcal{I} \right) d \geq 0$, for all $d \in \mathbb{R}^n$ with $\nabla h_i(x)^\top d = 0, i = 1, \ldots, m$,

(v) $|c_i(x)^2 \theta_i| \leq \varepsilon$ for all $i = 1, \ldots, p$.

Note that our first- and second-order optimality conditions given by Theorem 1 can be equivalently stated as, for all $\varepsilon > 0$, there exist ε-KKT and, respectively, ε-KKT2 points, arbitrarily close to x^*.

The first-order optimality condition is the generalization of the ones from [3, 8] to non-differentiable problems. In the smooth case, it implies the canonical first-order KKT conditions under weak constraint qualifications (see [5, 6, 7]), in particular, under linear constraints. The second-order optimality condition is the generalization of the one from [4, 37] to the non-differentiable case and it implies the canonical second-order KKT conditions defined in terms of the critical subspace under weak constraint qualifications, in particular, under linear constraints. When the constraints are smooth, a formulation of the optimality condition in terms of perturbed critical directions is presented in [15]. We note that the results from [37] can also be generalized without assuming smoothness on the boundary of C. In particular, without proving feasibility of the sequence $\{x^k\}$, the constant rank assumption can be dropped.

2.2 Sufficient Conditions for ε-Perturbed Stationary Points

Let us now focus on a special case of (4), where we assume $h(x) := \mathbf{A}x - \mathbf{b}$ and $c(x) := x$. This section then presents sufficient conditions for ε-KKT and ε-KKT2 points as per Definitions 1 and 2.

Proposition 1. *Given $\varepsilon > 0$, a sufficient condition for a point $x \in \mathbb{R}^n$ to be an ε-KKT point for problem* (1) *is the existence of $\lambda \in \mathbb{R}^m$ such that:*

(a) $\mathbf{A}x = \mathbf{b}, x > 0$,

(b) $\nabla f(x) + \mathbf{A}^\top \lambda \geq -\varepsilon$,

(c) $\|X(\nabla f(x) + \mathbf{A}^\top \lambda)\|_\infty \leq \varepsilon$.

Proof. Define $s := \max\{0, \nabla f(x) + \mathbf{A}^\top \lambda\}$ in Definition 1 and the claimed result follows from an easy calculation. □

Proposition 2. *Given $\varepsilon > 0$, a sufficient condition for a point $x \in \mathbb{R}^n$ to be an ε-KKT2 point for problem* (1) *is the existence of $\lambda \in \mathbb{R}^m$ such that:*

(a) $\mathbf{A}x = \mathbf{b}, x > 0$,

(b) $\nabla f(x) + \mathbf{A}^\top \lambda \geq -\varepsilon$,

(c) $\|X(\nabla f(x) + \mathbf{A}^\top \lambda)\|_\infty \leq \varepsilon$,

(d) $d^\top (X\nabla^2 f(x)X + \varepsilon\mathcal{I})d \geq 0$ *for all d such that $AXd = 0$.*

Proof. The claimed satisfaction of (i)-(iii) in Definition 2 follow immediately from Proposition 1. The following shows (iv) and (v). For all $\varepsilon' > 0$ it holds that $d^\top (X\nabla^2 f(x)X + (\varepsilon+\varepsilon')\mathcal{I})d > 0$ for all $d \neq 0$ such that $AXd = 0$. It is well know that, in this case, there is some $\rho > 0$ such that $X\nabla^2 f(x)X + (\varepsilon + \varepsilon')\mathcal{I} + \rho X\mathbf{A}^\top \mathbf{A}X$ is positive definite (see, for instance, [37, Proposition 2.1]). Since X^{-1} is positive definite, we have $\nabla^2 f(x) + \sum_{i=1}^m \frac{\varepsilon+\varepsilon'}{x_i^2} e_i e_i^\mathsf{T} + \rho\mathbf{A}^\top\mathbf{A}$ is positive definite, where e_i is the i-th canonical vector. Taking the limit $\varepsilon' \to 0^+$ and restricting to d with $\mathbf{A}d = 0$ we have $d^\top(\nabla^2 f(x) + \sum_{i=1}^m \frac{\varepsilon}{x_i^2} e_i e_i^\mathsf{T})d \geq 0$ for all d with $\mathbf{A}d = 0$ and the result follows defining $\theta_i := \frac{\varepsilon}{x_i^2}, i = 1, \ldots, n$. □

3 Interior Trust-Region Point Algorithms and Computational Complexity for ε-Perturbed Stationary Points

We once again focus on (1) and present two interior trust-region point (ITRP) algorithms that are theoretically ensured to generate ε-perturbed stationary points. Both algorithms belong to the class of fully polynomial time approximation schemes. Let $\Omega := \{x \mid \mathbf{A}x = \mathbf{b}, x \geq 0\}$ denote the feasible set and

$\Omega^\circ := \{x \mid \mathbf{A}x = \mathbf{b}, x > 0\}$ its interior. Assume that the feasible region is bounded and has a non-empty interior. For any given positive $\mu \leq 1$, we consider the potential function

$$\phi(x) := f(x) - \mu \sum_{i=1}^{n} \log(x_i). \tag{7}$$

Note that the gradient of the potential function at $x > 0$ is

$$\nabla\phi(x) = \nabla f(x) - \mu X^{-1} e.$$

Then the ITRP algorithms are summarized in Algorithm 1, where we have a specific initialization rule; we elect to initialize the algorithm with an approximate analytic center $x^0 \in \Omega^\circ$ that satisfies

$$-\sum_{i=1}^{n} \log(x_i) \geq -\sum_{i=1}^{n} \log(x_i^0) - O(1), \tag{8}$$

for all $x := (x_i) \in \Omega^\circ$ for some problem-independent constant $O(1)$. Such an initial solution is efficiently computable.

Meanwhile, we choose to terminate the algorithm when the per-iteration improvement on the potential function is smaller than a certain threshold to be specified soon afterwards. Constants μ and β will also be defined later on.

Algorithm 1 Pseudo-code of the interior trust-region point (ITRP) algorithm

Step 1. Given $\varepsilon \in (0, 1]$ and choose $x^0 \in \Omega^\circ$ to be an approximate analytic center of the feasible region. Let $t := 0$.

Step 2. Solve the following problem

$$\min \begin{cases} \nabla\phi(x^t)^\top X_t d & \text{first-order ITRP} \\ \nabla\phi(x^t)^\top X_t d + \frac{1}{2} d^\top X_t \nabla^2 f(x^t) X_t d & \text{second-order ITRP} \end{cases} \tag{9}$$

$$s.t.\ AX_t d = 0,\ \|d\| \leq \beta; \tag{10}$$

where $X_t = diag(x^t)$. Denote by d^t the solution.

Step 3. Update $x^{t+1} := x^t + X_t d^t$.

Step 4. Algorithm terminates if stopping criterion is satisfied. Otherwise, let $t := t + 1$ and go to Step 2.

In Algorithm 1, the per-iteration subproblem (9)-(10) can be chosen from the first-order or the second-order mode depending on the target of the optimization, that is, to achieve an ε-perturbed first- or second-order stationary point, respectively. Also, the second-order mode yields a perturbed first-order stationary point at a faster complexity rate. In both modes, the resulting per-iteration

problem (9)-(10) are easily solvable. Specifically, in the case of first-order ITRP, Problem (9)-(10) admits a closed form solution that does not involve any Hessian information, nor matrix inversion. Therefore, in this case the ITRP belongs to the class of first-order algorithms. In contrast, in the second-order ITRP, the subproblem can be solved using a bisection scheme as per [55, 54] with a "log-log" (quadratic) rate of complexity.

In the following, we will show that both modes of the ITRP entails the best rate of worst-case iteration complexity known for a stricter class of nonlinear optimization problems. We will make use of the following lemma, which is well known in the literature of interior-point algorithms (e.g., [40]):

Lemma 1. *Let $x > 0$ and $\|X^{-1}d\| \leq \beta < 1$. Then*

$$-\sum_{i=1}^{n} \ln(x_i + d_i) + \sum_{i=1}^{n} \ln(x_i) \leq -e^\top X^{-1} d + \frac{\beta^2}{2(1-\beta)}.$$

3.1 Complexity Analysis for the First-Order ITRP Algorithm

This subsection presents the complexity analysis for the first-order ITRP with a general assumption that f is potentially not (directionally) differentiable. In the following, we first present our assumptions in Section 3.1.1. Section 3.1.2 then presents the promised complexity analyses.

3.1.1 Assumptions for the first-order ITRP

Our complexity analysis herein relies on the following set of assumptions.

Assumption 3:

(a) Function $f(x)$ is differentiable for all $x \in \Omega^\circ$. In addition, there exists $\gamma \geq 1$ such that for all $x \in \Omega^\circ$ and $d \in \{d : \|d\| \leq r, X(e+d) \in \Omega\}$ for some $r < 1$,

$$f\left(X(e+d)\right) \leq f(x) + \langle X\nabla f(x), d\rangle + \frac{\gamma}{2}\|d\|^2.$$

(b) The feasible region is bounded with $\max\{\|x\|_\infty : x \in \Omega\} \leq R$, for some $R \geq 1$.

(c) The objective function is bounded from below in the feasible set, that is, there exists $L \in \mathbb{R}$ with $f(x) \geq L$ for all $x \in \Omega^\circ$.

Remark 1. Assumption 3.(a) subsumes the following special but important cases:

1. For all $x, x^+ \in \Omega$, it holds that $f(x^+) \leq f(x) + \langle \nabla f(x), x^+ - x \rangle + \frac{\hat{\beta}}{2}\|x^+ - x\|^2$ for some $\hat{\beta} > 0$. Such an inequality implies Assumption 3.(a) with $\gamma := \hat{\beta}R^2$.

2. Function $f := f_1 + f_2$ is a composite function, with f_1 being continuously differentiable and $f_2(x) := \sum_{i=1}^n x_i^p$ for any $p : 0 < p < 1$. To see this, we may observe that $f_2(X(d+e)) = \sum_{i=1}^n x_i^p (d_i+1)^p$ for any $d = (d_i) \in \mathbb{R}^n$ and any $x = (x_i) \in \Omega$. Also, $f_2(X(d+e))$ is continuously differentiable in d and the largest eigenvalue of its Hessian in d is upper bounded by $\frac{R^p p(p-1)}{(1-\beta)^{2-p}}$. It is worth noticing that f_2 is not differentiable when $x_i = 0$ for any i.

Remark 2. Assumption 3.(b) can be easily generalized to the assumption that the level sets of f are bounded, that is, given $x^0 \in \Omega^\circ$, there exists $R \geq 1$ such that $\sup\{\|x\|_\infty : f(x) \leq f(x^0), x \in \Omega^\circ\} \leq R$.

3.1.2 Complexity estimate for the first-order ITRP

We are now ready to present our complexity analysis. We elect to terminate the algorithm whenever $\phi(x^{t+1}) - \phi(x^t) > -\frac{\varepsilon^2}{2\gamma+4\varepsilon}$ and output the solution x^t.

Theorem 2. *Suppose that Assumption 3 holds. Denote by f^* the global minimal value of the objective function f on Ω. Consider Algorithm 1 with first-order ITRP per-iteration problem. For any $\varepsilon \in (0, \min\{r, 1\}]$, let $\mu := \varepsilon$, $\beta := (\gamma + 2\mu)^{-1}\mu$, and $t^* := \left\lceil \frac{\left(f(x^0) - f^* + O(1) - \varepsilon\right)(2\gamma+4\varepsilon)}{\varepsilon^2} \right\rceil$, the algorithm terminates before the t^*-th iteration at a 2ε-KKT point, more precisely, at a feasible solution $\hat{x}$ that satisfies $\nabla f(\hat{x}) + \mathbf{A}^\top \hat{y} > 0$ and $\|diag(\hat{x})\left(\nabla f(\hat{x}) + \mathbf{A}^\top \hat{y}\right)\|_\infty \leq 2\varepsilon$ for some $\hat{y}$. Otherwise, it holds that $f(x^{t^*}) - f^* \leq \varepsilon$.*

Proof. **Step 1.** In this step, we would like to show that $x^t \in \Omega^\circ$ for all $t \geq 1$. To this end, we notice that, if $x^{t-1} \in \Omega^\circ$, it holds that $x_i^t = x_i^{t-1} + x_i^{t-1} d_i^{t-1} = x_i^{t-1}(1 + d_i^{t-1}) > 0$ for any $i = 1, ..., n$, where the last inequality is because $\|d^{t-1}\| \leq \beta < 1$ imposed as a constraint in (10). Also, if $x^{t-1} \in \Omega^\circ$, it holds that $\mathbf{A}x^t = \mathbf{A}(x^{t-1} + X_{t-1}d^{t-1}) = \mathbf{b} + \mathbf{A}X_{t-1}d^{t-1} = \mathbf{b}$, where the last identity is based on constraint (10). Our proof for Step 1 completes by noticing that $x^0 \in \Omega^\circ$.

Step 2. In this step, we would like to show that either of the following holds at iteration k:

$$\phi(x^{t+1}) - \phi(x^t) \leq -\frac{\varepsilon^2}{2\gamma + 4\varepsilon}, \tag{11}$$

or $\|X_t \nabla f(x^t) - \mu e + X_t A^\top y^t\|_\infty < 2\varepsilon$ and $\nabla f(x^t) + \mathbf{A}^\top y^t > 0$ for some $y^t \in \mathbb{R}^m$.

To this end, we first notice that subproblem (9)-(10) can be solved globally, whose first-order optimality condition yields that

$$X_t \nabla f(x^t) - \mu e + X_t A^\top y^t + \lambda^t d^t = 0, \tag{12}$$

for some Lagrange multipliers $y^t \in \mathbb{R}^m$ and $\lambda^t \in \mathbb{R}$. From the inequality in Assumption 3.(a), since $x^t \in \Omega^\circ$ and $d^t : \|d^t\| \leq \beta = (\gamma + 2\mu)^{-1}\mu < \varepsilon \leq 1$ from the result in Step 1, it holds that

$$f\left(X_t(e + d^t)\right) \leq f(x^t) + \langle X_t \nabla f(x^t), d^t \rangle + \frac{\gamma}{2}\|d^t\|^2. \tag{13}$$

Combined with Lemma 1, it implies that

$$\begin{aligned} \phi(x^{t+1}) - \phi(x^t) \leq & \langle \nabla f(x^t),\, X_t d^t \rangle + \frac{\gamma}{2}\|d^t\|^2 - \mu e^\top X_t^{-1} d + \mu\beta^2 & (14) \\ = & \langle \nabla \phi(x^t),\, X_t d^t \rangle + \frac{\gamma}{2}\|d^t\|^2 + \mu\beta^2. & (15) \end{aligned}$$

Thus,

$$\phi(x^{t+1}) - \phi(x^t) \leq \langle X_t A^\top y^t - \lambda d^t,\, d^t \rangle + \frac{\gamma}{2}\|d^t\|^2 + \mu\beta^2 = \langle -\lambda^t d^t,\, d^t \rangle + \frac{\gamma}{2}\|d^t\|^2 + \mu\beta^2. \tag{16}$$

Case 1: If $\|d^t\| < \beta$, then $\lambda^t = 0$ and $X_t \nabla f(x^t) + X_t A^\top y^t = \mu e$. Since $\mu := \varepsilon > 0$, it therefore holds that $\nabla f(x^t) + \mathbf{A}^\top y^t > 0$ and that $\|X_t \nabla f(x^t) + X_t A^\top y^t\|_\infty \leq \varepsilon$.

Case 2: Consider the case where $\|d^t\| = \beta$. Let $p(x, y) := X\nabla f(x) - \mu e + XA^\top y$. (Again, $X := diag(x)$.) From (12), it therefore holds that $\|p(x^t, y^t)\| = \lambda^t \|d^t\| = \lambda^t \beta$. Combined with (16), it yields that

$$\phi(x^{t+1}) - \phi(x^t) \leq -\lambda^t \beta^2 + \frac{\gamma}{2}\|d^t\|^2 + \mu\beta^2 = -\beta\|p(x^t, y^t)\| + \left(\frac{\gamma}{2} + \mu\right)\beta^2. \tag{17}$$

Case 2.1: Under Case 2, if $\|p(x^t, y^t)\| \geq \mu$, then

$$\phi(x^{t+1}) - \phi(x^t) \leq -\beta\mu + \left(\frac{\gamma}{2} + \mu\right)\beta^2. \tag{18}$$

Since $\mu := \varepsilon$ and $\beta := (\gamma + 2\mu)^{-1}\mu$, we have that

$$\phi(x^{t+1}) - \phi(x^t) \leq -\frac{\varepsilon^2}{2\gamma + 4\varepsilon}. \tag{19}$$

Case 2.2: Under Case 2, if $\|p(x^t, y^t)\| < \mu$, then

$$\|X_t \nabla f(x^t) - \mu e + X_t A^\top y^t\|_\infty \leq \|X_t \nabla f(x^t) - \mu e + X_t A^\top y^t\| < \mu, \tag{20}$$

therefore, $X_t \nabla f(x^t) + X_t A^\top y^t > 0 \implies \nabla f(x^t) + \mathbf{A}^\top y^t > 0$. Meanwhile, $\|X_t \nabla f(x^t) + X_t A^\top y^t\|_\infty < 2\mu = 2\varepsilon$ for given $\mu := \varepsilon$. Summarizing the above cases, we know that Case 1, Case 2.1, and Case 2.2 are mutually exclusive. Thus we have the desired result in Step 2.

Step 3. We would like to summarize the above steps to obtain the claimed results in this theorem. We first observe that, because the elected initial solution x^0 satisfies that

$$-\sum_{i=1}^{n} \log(x_i^t) \geq -\sum_{i=1}^{n} \log(x_i^0) - O(1),$$

we have that, if (19) holds for all $t \leq t'$, it holds that

$$f(x^{t'}) - f(x^0) \leq -\frac{t'\varepsilon^2}{2\gamma + 4\varepsilon} + O(1). \tag{21}$$

It therefore holds that $f(x^{t'}) - f^* \leq \big[f(x^0) - f^*\big] - \frac{t'\varepsilon^2}{2\gamma+4\varepsilon} + O(1)$.

Recall that the algorithm terminates whenever $\phi(x^{t+1}) - \phi(x^t) > -\frac{\varepsilon^2}{2\gamma+4\varepsilon}$ for some t. Therefore, at iteration $t^* = \frac{\big(f(x^0)-f^*+O(1)-\varepsilon\big)(2\gamma+4\varepsilon)}{\varepsilon^2}$, it holds either that the algorithm has terminated before iteration k^* at a feasible solution $\hat{x}$ that satisfies that $\nabla f(\hat{x}) + \mathbf{A}^\top \hat{y} > 0$ and $\|diag(\hat{x})\nabla f(\hat{x}) + \hat{X}A^\top \hat{y}\|_\infty \leq \varepsilon$. Otherwise, it holds that $f(x^{k^*}) - f^* \leq \varepsilon$. □

Remark 3. The first-order ITRP solves a constrained problem with potential non-differentiability at an iteration complexity of $O(1/\varepsilon^2)$. For this types of problems, such a rate is best known to the literature. It is also worth emphasizing that the per-iteration problem admits a closed-form solution.

3.2 Complexity Analysis for the Second-Order ITRP Algorithm

This subsection presents the complexity analysis for the second-order ITRP with three different sets of regularities on f: (i) f is potentially not twice differentiable; (ii) f is potentially not differentiable; and (iii) f is a quadratic function. The resulting complexity estimates as well as the characteristics of the final solution output from the ITRP vary according to the changes of assumptions. In the following, we first present our assumptions in Section 3.2.1. Section 3.2.2 then presents the promised complexity analyses.

3.2.1 Assumptions for the second-order ITRP

The analysis on the second-order ITRP relies on the following assumptions.
Assumption 4: Function $f(x)$ is twice differentiable for all $x \in \Omega^\circ$. For all $x \in \Omega^\circ$ and $d, d' \in \{d : \|d\| \leq r,\, X(e+d) \in \Omega^\circ\}$, for some $r < 1$ and $\eta \geq 1$, it holds that

$$\|X\nabla^2 f\left(X(e+d)\right) - X\nabla^2 f\left(X(e+d')\right)\| \leq \eta\|d - d'\|; \quad \text{and}$$
$$\nabla f\left(X(e+d)\right) - \nabla f(x) \leq \langle X\nabla f(x),\, d\rangle + \frac{1}{2}d^\top X\nabla^2 f\left(x\right)Xd + \frac{\eta}{3}\|d\|^3. \tag{22}$$

Assumption 5: Function $f(x)$ is twice differentiable for all $x \in \Omega^\circ$. For all $x \in \Omega^\circ$ and $d, d' \in \{d : \|d\| \leq r,\, X(e+d) \in \Omega^\circ\}$, for some $r < 1$ and $\eta \geq 1$, it holds that

$$\|X\nabla^2 f\left(X(e+d)\right)X - X\nabla^2 f\left(X(e+d')X\right)\| \leq \eta\|d - d'\|; \quad \text{and}$$
$$\nabla f\left(X(e+d)\right) - \nabla f(x) \leq \langle X\nabla f(x),\, d\rangle + \frac{1}{2}d^\top X\nabla^2 f\left(x\right)Xd + \frac{\eta}{3}\|d\|^3. \tag{23}$$

Remark 4. Assumption 4 and Assumption 5 subsume some special but important cases:

1. For all $x, x^+ \in \Omega$, it holds that $f(x)$ is twice differentiable and

$$\|\nabla^2 f(x) - \nabla^2 f(x^+)\| \leq \hat{\eta}\|x - x^+\|, \tag{24}$$

for some $\hat{\eta} > 0$. Such an inequality implies both Assumptions 4 and Assumption 5 with $\eta := \hat{\eta}R^3$. These are immediate from the observation that

$$\|X\nabla^2 f(x)X - X\nabla^2 f(x^+)X\| \leq \|X\|^2\hat{\eta}\|x - x^+\| \leq \|X\|^3\hat{\eta}\|d\|, \tag{25}$$
$$\|X\nabla^2 f(x) - X\nabla^2 f(x^+)\| \leq \|X\|\hat{\eta}\|x - x^+\| \leq \|X\|^2\hat{\eta}\|d\|, \tag{26}$$

as well as the direct implication of (24) in the form of

$$\begin{aligned}\nabla f\left(X(e+d)\right) - \nabla f(x) \leq & \langle X\nabla f(x),\, d\rangle + \frac{1}{2}d^\top X\nabla^2 f\left(x\right)Xd + \frac{\hat{\eta}}{3}\|Xd\|^3 \\ \leq & \langle X\nabla f(x),\, d\rangle + \frac{1}{2}d^\top X\nabla^2 f\left(x\right)Xd + \frac{R^3\hat{\eta}}{3}\|d\|^3.\end{aligned}$$

2. Let function $f := f_1 + f_2$ be a composite function, with f_1 being twice continuously differentiable. If $f_2(x) := \sum_{i=1}^n x_i^p$ for some $p : p > 0$ then for any $d = (d_i) \in \mathbb{R}^n : \|d\| \leq r < 1$, we immediately have

$$\begin{aligned}\frac{\partial^2 f_2\left(X(d+e)\right)}{\partial x_i^2} &= p(p-1)x_i^{p-2}(d_i+1)^{p-2}; \\ x_i \cdot \frac{\partial^2 f_2\left(X(d+e)\right)}{\partial x_i^2} &= p(p-1)x_i^{p-1}(d_i+1)^{p-2}; \\ (x_i)^2 \cdot \frac{\partial^2 f_2\left(X(d+e)\right)}{\partial x_i^2} &= p(p-1)x_i^{p}(d_i+1)^{p-2}.\end{aligned}$$

Then, it is easily verifiable that:

- if $p : 1 < p < 2$, Assumption 4 holds, but $f(x)$ is not twice differentiable for $x \in \{x_i = 0, \text{ for some } i\}$.
- if $p : 0 < p < 1$, Assumption 5 holds, but $f(x)$ is not differentiable for $x \in \{x_i = 0, \text{ for some } i\}$.

Remark 5. Assumption 5 subsumes Assumption 4: It is evident that Assumption 4 implies Assumption 5, while the reverse does not hold telling from the second special case in Remark 4.

Assumption 6: f is a quadratic function, that is, $\eta = 0$.

3.2.2 Complexity estimates for the second-order ITRP

This section presents the complexity estimates for the second-order ITRP under three different sets of assumptions. Theorem 3 first considers the case when f is potentially not twice differentiable and shows that the desired ε-perturbed first- and second-order stationary point can be achieved with a rate of $O(\varepsilon^{-3/2})$ and $O(\varepsilon^{-3})$, respectively. Then, Theorem 4 generalizes to the case where f is potentially not (directionally) differentiable and shows that the same set of efficiency rates can be achieved in generating a weaker version of the ε-perturbed first- and second-order stationary point. Such a version of approximate necessary conditions is also studied by [13]. Finally, Theorem 5 presents a special case where f is a quadratic function. In such a case, the second-order ITRP is especially efficient and achieves the ε-perturbed first- and second-order stationary point both at rate of $O(\varepsilon^{-1})$. Theorem 5 presents an alternative proof for the same result presented in [55]. We should note that the termination criteria for the above three cases are slightly different.

For our first case, we consider the algorithm under Assumptions 4 and 6. We elect to terminate the second-order ITRP whenever the following criteria hold:

$$\phi(x^{t+1}) - \phi(x^t) > -\frac{\sqrt{\varepsilon^3}}{200\eta^2 R^{3/2}},$$
$$\phi(x^{t+2}) - \phi(x^{t+1}) > -\frac{\sqrt{\varepsilon^3}}{200\eta^2 R^{3/2}}.$$

At termination, the algorithm outputs solution x^{t+1}.

Theorem 3. *Suppose that Assumptions 3.(b), 3.(c) and 4 hold. Denote by f^* the global minimal value of the objective function f on Ω. Consider Algorithm 1 with second-order ITRP per-iteration problem. For any $\varepsilon \in \left(0, \min\left\{10\eta^2 r^2, \frac{1}{2}\right\}\right]$, let $\mu := \frac{\varepsilon}{5\eta R}$, $\beta := \mu^{1/2}\eta^{-1/2}/\sqrt{2}$, and $t^* := \left\lceil \frac{400\eta^2 R^{3/2}\left(f(x^0) - f^* + O(1)\varepsilon\right)(2\eta + 4\varepsilon)}{\sqrt{\varepsilon^3}} + 1 \right\rceil$. The algorithm terminates before the t^*-th iteration at an ε-KKT and $\sqrt{\varepsilon}$-KKT2 point, more precisely, at a feasible solution $\hat{x}$ that satisfies, for some $\hat{y} \in \mathbb{R}^m$, that*

$$\begin{aligned}
&\hat{x} > 0, \quad \nabla f(\hat{x}) + \mathbf{A}^\top \hat{y} > -\varepsilon,\\
&\|diag(\hat{x})(\nabla f(\hat{x}) + \mathbf{A}^\top \hat{y})\|_\infty \le \varepsilon,\\
&d^\top \left(diag(\hat{x})\nabla^2 f(\hat{x}) diag(\hat{x}) + \sqrt{\varepsilon} I\right) d \ge 0, \quad \forall d:\ \mathbf{A} diag(\hat{x}) d = 0.
\end{aligned}$$

Otherwise, it holds that $f(x^{t^*}) - f^* \le \varepsilon$.

Proof. **Step 1.** Following Step 1 of the proof for Theorem 2, it is straightforward that $x^t \in \Omega^\circ$ for all $t \ge 1$.

Step 2. We would like to show that if $\phi(x^{t+1}) - \phi(x^t) > -\frac{\sqrt{2\eta\mu^3}}{24\eta}$ then $\nabla^2 f(x^t) X_t d^t - \mathbf{A}^\top y^t + \nabla f(x^t) > 0$ and $0 \le x_i(\nabla f(x^t) + \nabla^2 f(x^t) d^t - A^\top y^t)_i \le 2\mu,\ \forall i$, for $\beta := \mu^{1/2}\eta^{-1/2}/\sqrt{2}$ and some $y^t \in \mathbb{R}^m$.

To this end, combine Assumption 4 with both $\|d^t\| \leq \beta = \mu^{1/2}\eta^{-1/2}/\sqrt{2} \leq r$ and Lemma 1. It therefore holds that

$$\begin{aligned}
&\phi(x^{t+1}) - \phi(x^t) \\
&\leq \nabla f(x^t)^\top X_t d^t + \frac{1}{2}(d^t)^\top X_t \nabla^2 f(x^t) X_t d^t + \frac{\eta}{3}\|d^t\|^3 - \mu e^\top X_t^{-1} d^t + \mu\beta^2 \\
&= \nabla \phi(x^t)^\top X_t d^t + \frac{1}{2}(d^t)^\top X_t \nabla^2 f(x^t) X_t d^t + \frac{\eta}{3}\|d^t\|^3 + \mu\beta^2 \\
&\leq \nabla \phi(x^t)^\top X_t d^t + \frac{1}{2}(d^t)^\top X_t \nabla^2 f(x^t) X_t d^t + \left(\frac{\eta}{3}\beta + \mu\right)\beta^2. \qquad (27)
\end{aligned}$$

Then, the necessary and sufficient global optimality conditions of the trust-region subproblem, besides the feasibility of d^t, are

$$\begin{aligned}
&(X_t \nabla^2 f(x^t) X_t + \lambda^t I) d^t - X_t \mathbf{A}^\top y^t = -X_t \nabla \phi(x^t); \\
&(X_t \nabla^2 f(x^t) X_t + \lambda^t I)_{AX_t} \succeq 0, \quad \lambda^t \geq 0, \quad \lambda^t(\beta - \|d^t\|) = 0;
\end{aligned} \qquad (28)$$

for Lagrange multipliers $y^t \in \mathbb{R}^m$ and $\lambda^t \in \mathbb{R}$, see [51, 50, 33]. Here, $(X_t \nabla^2 f(x^t) X_t + \lambda^t I)_{AX_t} \succeq 0$ means

$$d^\top (X_t \nabla^2 f(x^t) X_t + \lambda^t I) d \geq 0, \ \forall d \in \{d : \ AX_t d = 0\}.$$

If $\|d'\| = \beta$, let vector

$$p(x^t, y^t) = X_t \nabla^2 f(x^t) X_t d^t - X_t \mathbf{A}^\top y^t + X_t \nabla \phi(x^t).$$

Then from (28), we have

$$\lambda^t d^t = -p(x^t, y^t). \qquad (29)$$

Thus,

$$\begin{aligned}
&\nabla \phi(x^t)^\top X_t d^t + \frac{1}{2}(d^t)^\top X_t \nabla^2 f(x^t) X_t d^t \\
=&\frac{1}{2}\nabla \phi(x^t)^\top X_t d^t + \frac{1}{2}(d^t)^\top (X_t \nabla \phi(x^t) + X_t \nabla^2 f(x^t) X_t d^t) \\
=&\frac{1}{2}(\nabla \phi(x^t)^\top X_t - A^\top y^t)^\top d^t + \frac{1}{2}(d^t)^\top (X_t \nabla \phi(x^t) + X_t \nabla^2 f(x^t) X_t d^t - A^\top y) \\
=&-\frac{1}{2}(d^t)^\top (X_t \nabla^2 f(x^t) X_t + \lambda^t I) d^t + \frac{1}{2}(d^t)^\top p(x^t, y^t) \\
\leq&\frac{1}{2}(d^t)^\top p(x^t, y^t) = -\frac{1}{2}\lambda^t \|d^t\|^2, \qquad (30)
\end{aligned}$$

where (30) is immediately due to (29).

As an immediate result, combined with (27), it holds that

$$\begin{aligned}
\phi(x^{t+1}) - \phi(x^t) &\leq -\frac{1}{2}\lambda^t \|d^t\|^2 + \left(\frac{\eta}{3}\beta + \mu\right)\beta^2 \qquad (31) \\
&= -\frac{1}{2}\lambda^t \|d^t\|^2 + \left(\frac{\eta}{3}\mu^{1/2}\eta^{-1/2}/\sqrt{2} + \mu\right)\mu\eta^{-1}/2 \qquad (32) \\
&= -\frac{1}{2}\lambda^t \|d^t\|^2 + \left(\frac{\sqrt{2\eta\mu^3}}{12\eta} + \frac{\mu^2}{2\eta}\right). \qquad (33)
\end{aligned}$$

Recall that $\eta \geq 1$ and $\varepsilon \leq \frac{1}{2} \leq \frac{5\eta R^2}{2} \Longrightarrow \mu \leq \frac{\eta}{8} \Longrightarrow \frac{\sqrt{2\eta\mu^3}}{12\eta} + \frac{\mu^2}{2\eta} \leq \frac{5\sqrt{2\eta\mu^3}}{24\eta}$. If $\phi(x^{t+1}) - \phi(x^t) > -\frac{\sqrt{2\eta\mu^3}}{24\eta}$, then $-\frac{\sqrt{2\eta\mu^3}}{24\eta} < -\frac{1}{2}\lambda^t\|d^t\|^2 + \left(\frac{\eta}{3}\beta + \mu\right)\beta^2 \Longrightarrow \frac{1}{2}\lambda^t\|d^t\|^2 < \frac{\sqrt{2\eta\mu^3}}{4\eta}$. We might consider the following two cases.

Case 1. If $\|d^t\| < \beta$, it then holds that $\lambda^t = 0$. As a result, condition (28) yields that

$$X_t\nabla^2 f(x^t)X_t d^t - X_t\mathbf{A}^\top y^t + X_t\nabla\phi(x^t) = 0; \quad (X_t\nabla^2 f(x^t)X_t)_{AX_t} \succeq 0. \tag{34}$$

Thus, it holds that

$$\|X_t\nabla^2 f(x^t)X_t d^t - X_t\mathbf{A}^\top y^t + X_t\nabla f(x^t)\|_\infty = \mu < 2\mu, \tag{35}$$

and

$$\nabla^2 f(x^t)X_t d^t - \mathbf{A}^\top y^t + \nabla f(x^t) > 0. \tag{36}$$

Case 2. If $\|d^t\| = \beta$, then $\|p(x^t, y^t)\| = \lambda^t\beta$. Thus

$$\frac{\sqrt{2\eta\mu^3}}{4\eta} > \frac{1}{2}\lambda^t\|d^t\|^2 = \frac{1}{2}\lambda^t\beta^2 = \frac{1}{2}\beta\|p(x^t, y^t)\| = \frac{\sqrt{2\eta\mu}}{4}\|p(x^t, y^t)\|,$$

which means that $\|p(x^t, y^t)\| < \mu$, that is,

$$\begin{aligned}\mu >& \|X_t\nabla^2 f(x^t)X_t d^t - X_t\mathbf{A}^\top y^t + X_t\nabla\phi(x^t)\|_\infty \\ =& \|(X_t\nabla^2 f(x^t)X_t d^t - X_t\mathbf{A}^\top y^t + X_t\nabla f(x^t)) - \mu e\|_\infty,\end{aligned}$$

which implies

$$\nabla^2 f(x^t)X_t d^t - \mathbf{A}^\top y^t + \nabla f(x^t) > 0,$$

and

$$0 \leq x_i(\nabla f(x^t) + \nabla^2 f(x^t)d^t - A^\top y^t)_i \leq 2\mu, \ \forall i.$$

Combining Cases 1 and 2, we have the desired result in Step 2.

Step 3. We would like to show that once it holds that

$$\begin{aligned}&\nabla^2 f(x^t)X_t d^t - \mathbf{A}^\top y^t + \nabla f(x^t) > 0; \\ \text{and} \quad &0 \leq x_i(\nabla f(x^t) + \nabla^2 f(x^t)d^t - A^\top y^t)_i \leq 2\mu, \ \forall i.\end{aligned} \tag{37}$$

then, it simultaneously holds that, for some $\hat{y} \in \mathbb{R}^m$:

$$\begin{aligned}\nabla f(x^{t+1}) - \mathbf{A}^\top\hat{y} &> -\frac{\mu}{2} \\ |x_i^{t+1}(\nabla f(x^{t+1}) - A^\top\hat{y})_i| &\leq 4\mu + \mu R, \quad \forall i.\end{aligned} \tag{38}$$

To that end, notice that, since x^t, $x^{t+1} \in \Omega^\circ$, from mean value theorem, it holds that, for some $\tau \in [0, 1]$,

$$\begin{aligned}&\nabla f(x^{t+1}) - \nabla f(x^t) \\ =&\nabla^2 f(\tau(x^{t+1} - x^t) + x^t)(x^{t+1} - x^t) = \nabla^2 f(\tau(x^{t+1} - x^t) + x^t)X_t d^t,\end{aligned} \tag{39}$$

and thus

$$\begin{aligned}
&\|\nabla f(x^{t+1}) - \nabla f(x^t) - \nabla^2 f(x^t) X_t d^t\| \\
=&\| \left(\nabla^2 f(x^t) - \nabla^2 f(\tau(x^{t+1} - x^t) + x^t)\right) X_t d^t\| \\
=&\| \left(\nabla^2 f(X_t e) - \nabla^2 f(X_t(\tau d^t + e))\right) X_t\| \|d^t\| \\
\leq& \eta\tau \|d^t\|^2 \leq \eta \|d^t\|^2 \leq \eta\beta^2, \qquad (40)
\end{aligned}$$

where the last line is due to Assumption 4.(a), combined with $\|d\| \leq \beta < r$ and x^t, $x^{t+1} \in \Omega^\circ$, which will be useful soon afterwards.

Similarly, we also have

$$\begin{aligned}
&\|X_t \nabla f(x^{t+1}) - X_t \nabla f(x^t) - X_t \nabla^2 f(x^t) X_t d^t\| \\
=&\|X_t \left(\nabla^2 f(x^t) - \nabla^2 f(\tau(x^{t+1} - x^t) + x^t)\right) X_t d^t\| \\
\leq& \eta \|X_t\| \|d^t\|^2 \leq \eta R \|d^t\|^2 \leq \eta R \beta^2, \qquad (41)
\end{aligned}$$

Combining (37) with (40), we have that

$$\begin{aligned}
&\nabla f(x^{t+1}) - \mathbf{A}^\top y^t \\
\geq& \nabla^2 f(x^t) X_t d^t - \mathbf{A}^\top y^t + \nabla f(x^t) - \|\nabla f(x^{t+1}) - \nabla f(x^t) - \nabla^2 f(x^t) X_t d^t\|_\infty \\
\geq& -\eta\beta^2 = -\frac{\mu}{2}.
\end{aligned}$$

Meanwhile, combining (37) with (41), it obtains that

$$\begin{aligned}
&|x_i^{t+1}(\nabla f(x^{t+1}) - A^\top y^t)_i| \\
\leq& |(1 + d_i^t) x_i^t (\nabla f(x^t) + \nabla^2 f(x^t) d^t - A^\top y^t)_i| \\
&+ |1 + d_i^t| \cdot \|X_t \nabla f(x^{t+1}) - X_t \nabla f(x^t) - X_t \nabla^2 f(x^t) X_t d^t\|_\infty \\
\leq& (1+\beta)(2\mu + \eta R)\beta^2 \leq (1+\beta)\left(2\mu + \frac{\mu R}{2}\right) \leq 4\mu + \mu R.
\end{aligned}$$

The last line is due to $|1 + d_i^t| \leq (1 + \beta) \leq 2$.

Step 4. We would like to show that, if $\phi(x^{t+2}) - \phi(x^{t+1}) > -\frac{\sqrt{2\eta\mu^3}}{24\eta}$, then $(X_{t+1}\nabla^2 f(x^{t+1}) X_{t+1} + \sqrt{2\mu\eta} I)_{AX_{t+1}} \succeq 0$. To this end, we invoke (29) (where we let $t := t+1$), (57) (where we let $t := t+1$), and (28) (where we let $t := t+1$). The combination of the three results in

$$\left(X_{t+1}\nabla^2 f(x^{t+1}) X_{t+1} + \frac{\|p(x^{t+1}, y^{t+1})\|}{\beta} I\right)_{AX_{t+1}} \succeq 0. \qquad (42)$$

Further observe that from Step 2, it holds that, if $\phi(x^{t+2}) - \phi(x^{t+1}) > -\frac{\sqrt{2\eta\mu^3}}{24\eta}$, then $\frac{\|p(x^{t+1}, y^{t+1})\|}{\beta} \leq \frac{\mu}{\beta} = \sqrt{2\mu\eta}$. Combined with (63), we have the claimed result in this step.

Step 5. This step summarizes the above steps and prove the claimed results of the theorem.

We recall here x^0 is the approximate analytic center that satisfies

$$-\sum_{i=1}^{n}\log(x_i^t) \geq -\sum_{i=1}^{n}\log(x_i^0) - O(1), \tag{43}$$

where $O(1)$ is a constant.

We know that at iteration $t^* := \frac{400\eta^2 R^{3/2}\left(f(x^0)-f^*+O(1)-\varepsilon\right)(2\eta+4\varepsilon)}{\sqrt{\varepsilon^3}} + 1$, where $O(1)$ is the same number as in (52) if the termination criteria of simultaneously satisfying

$$\begin{aligned}\phi(x^{t+1}) - \phi(x^t) > -\frac{\sqrt{\varepsilon^3}}{200\eta^2 R^{3/2}} > -\frac{\sqrt{2\eta\mu^3}}{24\eta} = -\frac{\sqrt{10\varepsilon^3}}{600\eta^2 R^{3/2}},\\ \phi(x^{t+2}) - \phi(x^{t+1}) > -\frac{\sqrt{\varepsilon^3}}{200\eta^2 R^{3/2}},\end{aligned}$$

have never been satisfied, then, we obtain a reduction in the potential function:

$$\phi(x^{t*}) - \phi(x^0) \leq -\frac{\sqrt{\varepsilon^3}(t^*-1)}{400\eta^2 R^{3/2}} = -f(x^0) + f^* - O(1) + \varepsilon. \tag{44}$$

Then combined with (52), it holds that

$$\begin{aligned}f(x^{t*}) - f(x^0) - O(1) \leq -\frac{\sqrt{\varepsilon^3}(t^*-1)}{400\eta^2 R^{3/2}} = -f(x^0) + f^* - O(1) + \varepsilon\\ \Longrightarrow f(x^{t*}) - f^* \leq \varepsilon.\end{aligned} \tag{45}$$

Otherwise, the algorithm terminates before t^* and achieves a solution that satisfies

$$\begin{aligned}\nabla f(x^{t+1}) - \mathbf{A}^\top \hat{y} > -\frac{\mu}{2} > -\varepsilon,\\ |x_i^{t+1}(\nabla f(x^{t+1}) - A^\top \hat{y})_i| \leq 4\mu + \mu R \leq \varepsilon, \quad \forall i,\end{aligned} \tag{46}$$

according to Step 2. Furthermore, from Step 4, the satisfaction of the termination criteria also implies

$$\begin{aligned}\left(X_{t+1}\nabla^2 f(x^{t+1})X_{t+1} + \sqrt{2\mu\eta}I\right)_{AX_{t+1}} \succeq 0\\ \Longrightarrow \left(X_{t+1}\nabla^2 f(x^{t+1})X_{t+1} + \sqrt{\varepsilon}I\right)_{AX_{t+1}} \succeq 0,\end{aligned}$$

thus immediately leads to the desired result. □

Consider the same algorithm procedure as in the second-order ITRP. If the regularity on f is relaxed from Assumption 4 to Assumption 5, then we may still obtain an approximate KKT condition. Nonetheless, such an approximation is in a critically weaker form. Specifically, we have the following theorem. In this

case, we have a slightly different termination criterion: we elect to terminate the second-order ITRP whenever the following criteria hold:

$$\phi(x^{t+1}) - \phi(x^t) > -\frac{\sqrt{\varepsilon^3}}{200\eta^2},$$
$$\phi(x^{t+2}) - \phi(x^{t+1}) > -\frac{\sqrt{\varepsilon^3}}{200\eta^2}.$$

Once the algorithm terminates, it outputs x^{t+2} as our final solution.

Theorem 4. *Suppose that Assumptions 3.(b) and 3.(c) and 5 hold. Denote by f^* the global minimal value of the objective function f on Ω. Consider Algorithm 1 with second-order ITRP per-iteration problem. For any $\varepsilon \in \left(0,\, \min\left\{10\eta^2 r^2,\, \frac{1}{2}\right\}\right]$, let $\mu := \frac{\varepsilon}{5\eta}$, $\beta := \mu^{1/2}\eta^{-1/2}/\sqrt{2}$, and $t^* := \left\lceil \frac{400\eta^2\left(f(x^0)-f^*+O(1)-\varepsilon\right)(2\eta+4\varepsilon)}{\sqrt{\varepsilon^3}} + 1 \right\rceil$. The algorithm terminates before the t^*-th iteration at a feasible solution $\hat{x}$ that satisfies that*

$$\begin{aligned} &\hat{x} > 0, \quad \|diag(\hat{x})(\nabla f(\hat{x}) + \mathbf{A}^\top \hat{y})\|_\infty \le \varepsilon, \\ &d^\top \left(diag(\hat{x})\nabla^2 f(\hat{x}) diag(\hat{x}) + \sqrt{\varepsilon} I\right) d \ge 0, \quad \forall d:\ \mathbf{A} diag(\hat{x}) d = 0. \end{aligned} \tag{47}$$

Otherwise, it holds that $f(x^{t^*}) - f^* \le \varepsilon$.

Proof. **Step 1.** Following Step 1 of the proof for Theorem 2, it is straightforward that $x^t \in \Omega^\circ$ for all $t \ge 1$.

Step 2. Following Step 2 of the proof for Theorem 3, it is also evident that, if $\phi(x^{t+1}) - \phi(x^t) > -\frac{\sqrt{2\eta\mu^3}}{24\eta}$ then $0 \le x_i(\nabla f(x^t) + \nabla^2 f(x^t)d^t - A^\top y^t)_i \le 2\mu,\ \forall i$, for $\beta := \mu^{1/2}\eta^{-1/2}/\sqrt{2}$.

Step 3. We would like to show that once it holds that

$$0 \le x_i(\nabla f(x^t) + \nabla^2 f(x^t)d^t - A^\top y^t)_i \le 2\mu,\ \forall i. \tag{48}$$

then, it holds that, for some $\hat{y} \in \mathbb{R}^m$:

$$|x_i^{t+1}(\nabla f(x^{t+1}) - A^\top \hat{y})_i| \le 5\mu, \quad \forall i. \tag{49}$$

To that end, notice that, since $x^t,\ x^{t+1} \in \Omega^\circ$, from mean value theorem, it holds that, for some $\tau \in [0,\, 1]$,

$$\begin{aligned} &\nabla f(x^{t+1}) - \nabla f(x^t) = \nabla^2 f(\tau(x^{t+1} - x^t) + x^t)(x^{t+1} - x^t) \\ =&\nabla^2 f(\tau(x^{t+1} - x^t) + x^t) X_t d^t, \end{aligned}$$

and thus

$$\begin{aligned} &\|X_t \nabla f(x^{t+1}) - X_t \nabla f(x^t) - X_t \nabla^2 f(x^t) X_t d^t\| \\ =&\|X_t \left(\nabla^2 f(x^t) - \nabla^2 f(\tau(x^{t+1} - x^t) + x^t)\right) X_t d^t\| \\ =&\|X_t \left(\nabla^2 f(X_t e) - \nabla^2 f(X_t(\tau d^t + e))\right) X_t\| \|d^t\| \\ \le&\eta\tau\|d^t\|^2 \le \eta\|d^t\|^2 \le \eta\beta^2, \end{aligned} \tag{50}$$

where the last line is due to Assumption 5, combined with $\|d\| \leq \beta < r$ and x^t, $x^{t+1} \in \Omega^\circ$, which will be useful soon afterwards.

Combining (37) with (50), we have that

$$\begin{aligned}
&\nabla f(x^{t+1}) - \mathbf{A}^\top y^t \\
\geq &\nabla^2 f(x^t) X_t d^t - \mathbf{A}^\top y^t + \nabla f(x^t) - \|\nabla f(x^{t+1}) - \nabla f(x^t) - \nabla^2 f(x^t) X_t d^t\|_\infty \\
\geq &-\eta\beta^2 = -\frac{\mu}{2}.
\end{aligned}$$

Meanwhile, combining (48) with (50), it obtains that

$$\begin{aligned}
&|x_i^{t+1}(\nabla f(x^{t+1}) - A^\top y^t)_i| \\
\leq &|(1 + d_i^t) x_i^t (\nabla f(x^t) + \nabla^2 f(x^t) d^t - A^\top y^t)_i| \\
&+ |1 + d_i^t| \cdot \|X_t \nabla f(x^{t+1}) - X_t \nabla f(x^t) - X_t \nabla^2 f(x^t) X_t d^t\|_\infty \\
\leq &(1+\beta)(2\mu + \eta)\beta^2 \leq (1+\beta)\left(2\mu + \frac{\mu}{2}\right) \leq 5\mu.
\end{aligned}$$

The last line is due to $|1 + d_i^t| \leq (1 + \beta) \leq 2$.

Step 4. We would like to show that, if $\phi(x^{t+2}) - \phi(x^{t+1}) > -\frac{\sqrt{2\eta\mu^3}}{24\eta}$, then $(X_{t+1}\nabla^2 f(x^{t+1}) X_{t+1} + \sqrt{2\mu\eta} I)_{AX_{t+1}} \succeq 0$. To this end, we invoke (29) (where we let $t := t+1$), (57) (where we let $t := t+1$), and (28) (where we let $t := t+1$). The combination of the three results in

$$\left(X_{t+1}\nabla^2 f(x^{t+1}) X_{t+1} + \frac{\|p(x^{t+1}, y^{t+1})\|}{\beta} I\right)_{AX_{t+1}} \succeq 0. \tag{51}$$

Further observe that from Step 2, it holds that, if $\phi(x^{t+2}) - \phi(x^{t+1}) > -\frac{\sqrt{2\eta\mu^3}}{24\eta}$, then $\frac{\|p(x^{t+1}, y^{t+1})\|}{\beta} \leq \frac{\mu}{\beta} = \sqrt{2\mu\eta}$. Combined with (63), we have the claimed result in this step.

Step 5. This step summarizes the above steps and prove the claimed results of the theorem.

We recall here x^0 is the approximate analytic center that satisfies

$$-\sum_{i=1}^n \log(x_i^t) \geq -\sum_{i=1}^n \log(x_i^0) - O(1), \tag{52}$$

where $O(1)$ is a constant.

We know that at iteration $t^* = \frac{400\eta^2\left(f(x^0) - f^* + O(1) - \varepsilon\right)(2\eta + 4\varepsilon)}{\sqrt{\varepsilon^3}} + 1$, where $O(1)$ is the same number as in (52) if the termination criteria of simultaneously satisfying

$$\begin{aligned}
\phi(x^{t+1}) - \phi(x^t) > -\frac{\sqrt{\varepsilon^3}}{200\eta^2} > -\frac{\sqrt{2\eta\mu^3}}{24\eta} = -\frac{\sqrt{10\varepsilon^3}}{600\eta^2}, \\
\phi(x^{t+2}) - \phi(x^{t+1}) > -\frac{\sqrt{\varepsilon^3}}{200\eta^2},
\end{aligned}$$

have never been satisfied. Then, we obtain a reduction in the potential function:

$$\phi(x^{t*}) - \phi(x^0) \leq -\frac{\sqrt{\varepsilon^3}(t^* - 1)}{400\eta^2} = -f(x^0) + f^* - O(1) + \varepsilon. \tag{53}$$

Then combined with (52), it holds that

$$\begin{aligned} f(x^{t*}) - f(x^0) - O(1) \leq & -\frac{\sqrt{\varepsilon^3}(t^* - 1)}{400\eta^2} = -f(x^0) + f^* - O(1) + \varepsilon \\ \Longrightarrow & f(x^{t*}) - f^* \leq \varepsilon. \end{aligned}$$

Otherwise, the algorithm terminates before t^* and achieves a solution that satisfies

$$|x_i^{t+1}(\nabla f(x^{t+1}) - A^\top \hat{y})_i| \leq 5\mu \leq \varepsilon, \quad \forall i, \tag{54}$$

according to Step 2. Furthermore, from Step 4, the satisfaction of the termination criteria also implies

$$\begin{aligned} & \left(X_{t+1}\nabla^2 f(x^{t+1})X_{t+1} + \sqrt{2\mu\eta}I\right)_{AX_{t+1}} \succeq 0 \\ \Longrightarrow & \left(X_{t+1}\nabla^2 f(x^{t+1})X_{t+1} + \sqrt{\varepsilon}I\right)_{AX_{t+1}} \succeq 0, \end{aligned}$$

thus immediately leads to the desired result. □

Remark 6. We observe that even though (47) is a weaker condition than the desired one in this paper, it still applies to application problems such as the non-Lipschitz problem formulation of sparse optimization discussed by [13], who provide a different algorithm with the same complexity for a special case that satisfies all our assumptions.

We now consider a special case where substantially faster iteration complexity can be achieved. Such a result is, in fact, first presented by [55] for achieving an approximate first-order KKT point for linearly constrained nonconvex quadratic program. The complexity in the approximation to the second-order necessary condition has not been explicitly stated, though a closer look at the results therein may find it an immediate result from the paper. In the following, we provide an alternative proof for the complexity analysis, which results in some new insights in solving this type of problem. We elect to terminate the second-order ITRP whenever the following criteria hold:

$$\begin{aligned} \phi(x^{t+1}) - \phi(x^t) > -\frac{\varepsilon}{32}, \\ \phi(x^{t+2}) - \phi(x^{t+1}) > -\frac{\varepsilon}{32}. \end{aligned}$$

Once the algorithm terminates, it outputs x^{t+2} as our final solution.

Theorem 5. *Suppose that Assumptions 3.(b), 3.(c) and 6 hold. Denote by f^* the global minimal value of the objective function f on Ω. Consider Algorithm 1*

with second-order ITRP per-iteration problem. For any $\varepsilon \in \left(0,\, \min\left\{10\eta^2 r^2,\, \frac{1}{2}\right\}\right]$, *let* $\mu := \frac{\varepsilon}{4}$, $\beta := 1/4$, *and* $t^* := \left\lceil \frac{64(f(x^0)-f^*+O(1)-\varepsilon)+1}{\epsilon} \right\rceil$, *the algorithm terminates before the* t^**-th iteration at an* ε*-KKT2 point, more precisely, at a feasible solution* $\hat{x}$ *that satisfies that*

$$\begin{aligned} &\hat{x} > 0, \quad \nabla f(\hat{x}) - \mathbf{A}^\top \hat{y} > 0; \quad \|diag(\hat{x})(\nabla f(\hat{x}) + \mathbf{A}^\top \hat{y})\|_\infty \le \varepsilon, \\ &d^\top \left(diag(\hat{x})\nabla^2 f(\hat{x}) diag(\hat{x}) + \varepsilon I\right) d \ge 0, \quad \forall d:\ \mathbf{A} diag(\hat{x}) d = 0. \end{aligned} \tag{55}$$

Otherwise, it holds that $f(x^{t^*}) - f^* \le \varepsilon$.

Proof. **Step 1.** Following Step 1 of the proof for Theorem 2, it is straightforward that $x^t \in \Omega^\circ$ for all $t \ge 1$.

Step 2. We would like to show that if $\phi(x^{t+1}) - \phi(x^t) > -\frac{\mu}{16}$ then $0 \le x_i(\nabla f(x^t) + \nabla^2 f(x^t) d^t - A^\top y^t)_i \le 2\mu,\ \forall i$, for $\beta := 1/4$.

Following Step 2 of the proof for Theorem 3, while noticing that $\eta = 0$, we can show that it is also evident that,

$$\phi(x^{t+1}) - \phi(x^t) \le -\frac{1}{2}\lambda^t \|d^t\|^2 + \mu\beta^2. \tag{56}$$

Case 1. If $\|d^t\| < \beta$, it then holds that $\lambda^t = 0$. As a result, condition (28) yields that

$$X_t \nabla^2 f(x^t) X_t d^t - X_t \mathbf{A}^\top y^t + X_t \nabla \phi(x^t) = 0; \quad (X_t \nabla^2 f(x^t) X_t)_{AX_t} \succeq 0. \tag{57}$$

Thus, it holds that

$$\|X_t \nabla^2 f(x^t) X_t d^t - X_t \mathbf{A}^\top y^t + X_t \nabla f(x^t)\|_\infty = \mu < 2\mu, \tag{58}$$

and

$$\nabla^2 f(x^t) X_t d^t - \mathbf{A}^\top y^t + \nabla f(x^t) > 0. \tag{59}$$

Case 2. If $\|d^t\| = \beta$, then $\|p(x^t, y^t)\| = \lambda^t \beta$. Combined with $\mu\beta^2 = \frac{\mu}{16}$, it holds that

$$\frac{\mu}{8} > \frac{1}{2}\lambda^t \|d^t\|^2 = \frac{1}{2}\lambda^t \beta^2 = \frac{1}{2}\beta \|p(x^t, y^t)\| = \frac{1}{8}\|p(x^t, y^t)\|.$$

which means that $\|p(x^t, y^t)\| < \mu$, that is,

$$\begin{aligned} \mu >& \|X_t \nabla^2 f(x^t) X_t d^t - X_t \mathbf{A}^\top y^t + X_t \nabla \phi(x^t)\|_\infty \\ =& \|(X_t \nabla^2 f(x^t) X_t d^t - X_t \mathbf{A}^\top y^t + X_t \nabla f(x^t)) - \mu e\|_\infty, \end{aligned}$$

which implies

$$\nabla^2 f(x^t) X_t d^t - \mathbf{A}^\top y^t + \nabla f(x^t) > 0,$$

and

$$0 \le x_i(\nabla f(x^t) + \nabla^2 f(x^t) d^t - A^\top y^t)_i \le 2\mu,\ \forall i.$$

Combining Cases 1 and 2, we have the desired result in Step 2.

Step 3. We would like to show that once it holds that

$$\begin{aligned} &\nabla^2 f(x^t) X_t d^t - \mathbf{A}^\top y^t + \nabla f(x^t) > 0; \\ \text{and} \quad &0 \le x_i(\nabla f(x^t) + \nabla^2 f(x^t) d^t - A^\top y^t)_i \le 2\mu, \ \forall i, \end{aligned} \tag{60}$$

then, it simultaneously holds that, for some $\hat{y} \in \mathbb{R}^m$:

$$\begin{aligned} \nabla f(x^{t+1}) - \mathbf{A}^\top \hat{y} > 0, \\ |x_i^{t+1}(\nabla f(x^{t+1}) - A^\top \hat{y})_i| \le \mu, \quad \forall i. \end{aligned} \tag{61}$$

To that end, notice that, due to Assumption 7,

$$\nabla f(x^{t+1}) - \nabla f(x^t) = \nabla^2 f(x^t) X_t d^t. \tag{62}$$

Combining (60) with (62), we have that

$$\nabla f(x^{t+1}) - \mathbf{A}^\top y^t = \nabla^2 f(x^t) X_t d^t - \mathbf{A}^\top y^t + \nabla f(x^t) > 0.$$

Meanwhile, combining (60) with (62), it obtains that

$$\begin{aligned} &|x_i^{t+1}(\nabla f(x^{t+1}) - A^\top y^t)_i| \\ \le &|(1 + d_i^t) x_i^t (\nabla f(x^t) + \nabla^2 f(x^t) d^t - A^\top y^t)_i| \\ \le &2\mu(1+\beta)\beta^2 \le \mu. \end{aligned}$$

The last line is due to $|1 + d_i^t| \le (1 + \beta) \le 2$.

Step 4. We would like to show that, if $\phi(x^{t+2}) - \phi(x^{t+1}) > -\frac{\mu}{16}$, then $(X_{t+1}\nabla^2 f(x^{t+1}) X_{t+1} + 4\mu I)_{AX_{t+1}} \succeq 0$. To this end, we invoke (29) (where we let $t := t+1$), (57) (where we let $t := t+1$), and (28) (where we let $t := t+1$). The combination of the three results gives

$$\left(X_{t+1}\nabla^2 f(x^{t+1}) X_{t+1} + \frac{\|p(x^{t+1}, y^{t+1})\|}{\beta} I \right)_{AX_{t+1}} \succeq 0. \tag{63}$$

Further observe that from Step 2, it holds that, if $\phi(x^{t+2}) - \phi(x^{t+1}) > -\frac{\mu}{16}$, then $\frac{\|p(x^{t+1}, y^{t+1})\|}{\beta} \le \frac{\mu}{\beta} = 4\mu$. Combined with (63), we have the claimed result in this step. The rest of the proof is straightforward following Step 5 of the proof for Theorem 3, while we let $\mu := \frac{\varepsilon}{4}$ and $t^* := \frac{64(f(x^0) - f^* + O(1)\varepsilon) + 1}{\epsilon}$. □

Remark 7. We notice the substantial improvement in the iteration complexity: If f is quadratic, the complexity in achieving an ε-perturbed first-order and second-order stationary point is both $O(\varepsilon^{-1})$, while for the same algorithm to solve a more general problem, our complexity estimates are $O(\varepsilon^{-3/2})$ and $O(\varepsilon^{-3})$ for the first-order and second-order stationary points, respectively. The cause of this gap, to our understanding, is whether the cubic error term is present in the Taylor expansion-like inequalities (22) and (23), or namely, whether $\eta = 0$ holds. Note that when the p-th order derivative is used to find a first-order stationary

point with a more general set of convex constraints, the best known iteration complexity is $O(\varepsilon^{-(p+1)/p})$ [14, 23] (but with a costly per-iteration complexity). The quadratic case here discussed is compatible with this result as a limiting case $p \to +\infty$.

Remark 8. In all three cases of discussion above, the per-iteration problem of the second-order ITRP admits a bisection scheme as per [55, 54] with a "log-log" (quadratic) rate of complexity.

4 Conclusion

In this paper we consider the minimization of a continuous function that is potentially not differentiable or not twice-differentiable on the boundary of the feasible region. To characterize computable stationary points, we present suitable first- and second-order optimality conditions for this problem that generalizes to classical ones when the derivative on the boundary is available, through the use of an interior point technique. As a result, such an optimality condition is stronger than the existing conditions commonly used in the literature. We further develop new interior trust-region point algorithms and present their worst-case complexity estimates to solve the special but important case with linear constraints. Even with a weaker regularity on the objective function, the presented algorithms are theoretically guaranteed to yield a stronger optimality condition at the same best known complexity rates in the literature for first- and second-order stationarity using first- and second-order derivatives. We believe that this approach can be generalized for non-linear constraints and for infeasible initialization. Also, solving a higher-order subproblem, we believe this approach can yield iteration complexity results for finding q-th order stationary points, extending the results from [22].

Acknowledgement

This work was supported by the São Paulo Research Foundation (FAPESP grants 2013/05475-7 and 2016/02092-8) and the Brazilian National Council for Scientific and Technological Development (CNPq). The content is solely the responsibility of the authors and does not necessarily represent the official views of the FAPESP and CNPq.

References

[1] Audet, C., Dennis Jr., J.E.: Mesh adaptive direct search algorithms for constrained optimization. SIAM J. Optim. 17: 188-217 (2006)

[2] Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., Ma, T.: Finding local minima for nonconvex optimization in linear time. *arXiv:1611.01146*. (2016)

[3] Andreani, R., Haeser, G., Martinez, J. M.: On sequencial optimality conditions for smooth constrained optimization. *Optimization*, 60(5):627–641 (2011)

[4] Andreani, R., Haeser, G., Ramos, A., Silva, P. J. S.: A second-order sequential optimality condition associated to the convergence of optimization algorithms. *IMA Journal of Numerical Analysis*, DOI: 10.1093/imanum/drw064 (2017)

[5] Andreani, R., Haeser, G., Schuverdt, M. L., Silva, P. J. S.: A relaxed constant positive linear dependence constraint qualification and applications. *Mathematical Programming*, 135:255–273 (2012)

[6] Andreani, R., Martínez, J. M., Ramos, A., Silva, P. J. S.: Two new weak constraint qualifications and applications. *SIAM Journal on Optimization*, 22:1109–1135 (2012)

[7] Andreani, R., Martínez, J. M., Ramos, A., Silva, P. J. S.: A cone-continuity constraint qualification and algorithmic consequences. *SIAM Journal on Optimization*, 26(1):96–110 (2016)

[8] Andreani, R., Martinez, J. M., Svaiter, B. F.: A new sequencial optimality condition for constrained optimization and algorithmic consequences. *SIAM Journal of Optimization*, 20(6):3533–3554 (2010)

[9] Behling, R., Haeser, G., Ramos, A., Viana, D. S.: On a conjecture in second-order optimality conditions. *Optimization Online* (2016)

[10] Bian, W., Chen, X.: Optimality and complexity for constrained optimization problems with nonconvex regularization. *to appear in Mathematics of Operations Research* (2016)

[11] Bian, W., Chen, X.: Worst-case complexity of smoothing quadratic regularization methods for non-lipschitzian optimization. *SIAM Journal on Optimization*, 23(3):1718–1741 (2013)

[12] Bian, W., Chen, X.: Linearly constrained non-lipschitz optimization for image restoration. *SIAM Journal on Imaging Sciences*, 8(4):2294–2322 (2015)

[13] Bian, W., Chen, X., Ye, Y.: Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. *Mathematical Programming*, 149(1):301–327 (2015)

[14] Birgin, E. G., Gardenghi, J. L., Martínez, J. M., Santos, S. A., Toint, Ph. L.: Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, DOI: 10.1007/s10107-016-1065-8 (2016)

[15] Birgin, E. G., Haeser, G., Ramos, A.: Augmented lagrangians with constrained subproblems and convergence to second-order stationary points. *Optimization Online* (2016)

[16] Birgin, E. G., Martínez, J. M.: Quadratic regularization with cubic descent for unconstrained optimization. *Optimization Online*. (2016)

[17] Carmon, Y. J., Duchi, C., Hinder, O., Sidford, A.: Accelerated methods for nonconvex optimization. *ArXiv: 1611.00756*. (2016)

[18] Cartis, C., Gould, N. I. M., Toint, Ph. L.: Adaptive cubic regularisation methods for unconstrained optimization. part ii: worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, 130(2):295–319 (2011)

[19] Cartis, C., Gould, N. I. M., Toint, Ph. L.: On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739 (2011)

[20] Cartis, C., Gould, N. I. M., Toint, Ph. L.: On the complexity of finding first-order critical points in constrained nonlinear optimization. *Mathematical Programming*, 144(1):93–106 (2014)

[21] Cartis, C., Gould, N. I. M., Toint, Ph. L.: On the evaluation complexity of constrained nonlinear least-squares and general constrained nonlinear optimization using second-order methods. *SIAM Journal on Numerical Analysis*, 53(2):836–851 (2015)

[22] Cartis, C., Gould, N. I. M., Toint, Ph. L.: Second-order optimality and beyond: characterization and evaluation complexity in convexly-constrained nonlinear optimization. Report naXys-06-2016, Dept of Mathematics, UNamur, Namur (B). (2016)

[23] Cartis, C., Gould, N. I. M., Toint, Ph. L.: Toint. Evaluation complexity for smooth constrained optimization using scaled KKT conditions and high-order models. Online at: http://perso.fundp.ac.be/ phtoint/pubs/NTR-11-2015-R1.pdf (2015)

[24] Cartis, C., Gould, N. I. M., Toint, Ph. L.: Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, 28(1):93 – 108 (2012)

[25] Chen, X., Lu, Z., Pong, T. K.: Penalty methods for a class of non-lipschitz optimization problems. *SIAM Journal on Optimization*, 26(3):1465–1492 (2016)

[26] Chen, X., Xu, F., Ye, Y.: Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization. *SIAM Journal on Scientific Computing*, 32(5):2832–2852 (2010)

[27] Curtis, F. E., Robinson, D. P., Samadi, M.: A trust region algorithm with a worst-case iteration complexity of $O(\varepsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming*, DOI: 10.1007/s10107-016-1026-2 (2016)

[28] Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, (96):1348–1360 (2001)

[29] Fan, J., Lv, J.: Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory*, (57):5467–5484 (2011)

[30] Fan, J., Lv, J., Qi, L.: Sparse high dimensional models in economics. *Annu. Rev. Econom*, (3):291–317 (2011)

[31] Fan, J., Xue, L., Zou, H.: Strong oracle optimality of folded concave penalized estimation. *Ann. Stat.*, 3(42):819–849 (2014)

[32] Fiacco, A. V., McCormick, G. P.: *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Wiley (1968)

[33] Gay, D. M., Computing optimal locally constrained steps. *SIAM Journal on Scientific and Statistical Computing*, 2(2):186–197 (1981)

[34] Grapiglia, G. N., Yuan, J., Yuan, Y.: On the convergence and worst-case complexity of trust-region and regularization methods for unconstrained optimization. *Mathematical Programming*, 152(1):491–520 (2015)

[35] Grapiglia, G. N., Yuan, J., Yuan, Y.: Nonlinear stepsize control algorithms: Complexity bounds for first- and second-order optimality. *Journal of Optimization Theory and Applications*, 171(3):980–997 (2016)

[36] Gratton, S., Sartenaer, A., Toint, Ph. L.: Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, 19(1):414–444 (2008)

[37] Haeser, G.: A second-order optimality condition with first- and second-order complementarity associated to global convergence of algorithms. *Optimization Online* (2016)

[38] Han, S., Pool, J., Tran, J., Dally, W. J.: Learning both Weights and Connections for Efficient Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1135–1143, (2015)

[39] Jahn, J.: *Introduction to the Theory of Nonlinear Optimization*, Springer (2007)

[40] Karmarkar. N., A new polynomial-time algorithm for linear programming. *Combinatorica*, (4):373–395 (1984)

[41] Liu, H., Yao, T., Li, R., Ye, Y.: Folded concave penalized sparse linear regression: Sparsity, statistical performance, and algorithmic theory for local solutions, Online at: http://web.stanford.edu/ yyye/FoldedPenalty.pdf

[42] Liu, Y.-F., Ma, S., Dai, Y.-H., Zhang, S.: A smoothing SQP framework for a class of composite l_q minimization over polyhedron. *Mathematical Programming*, 158(1):467–500 (2016)

[43] Loh, P.-L., Wainwright, M. J.: Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.*, (16):559–616 (2015)

[44] Martínez, J. M., Raydan, M.: Cubic-regularization counterpart of a variable-norm trust-region method for unconstrained minimization. *Journal of Global Optimization* DOI: 10.1007/s10898-016-0475-8 (2016)

[45] Negahban, S. N., Ravikumar, P., Wainwright, M. J., B. Yu, B.: A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 4(27):538–557 (2012)

[46] Nesterov, Y.: *Introductory Lectures on Convex Optimization.* Springer Science+Business Media, LLC (2004)

[47] Nesterov, Y., Polyak, B. T.: Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205 (2006)

[48] Rockafellar, R. T., Wets, R. J. B.: *Variational Analysis.* Springer (1998)

[49] Toint, Ph. L.: Nonlinear stepsize control, trust regions and regularizations for unconstrained optimization. *Optimization Methods and Software*, 28(1):82–95 (2013)

[50] Sorensen, D. C.: Newton's method with a model trust region modification. *SIAM Journal on Numerical Analysis*, 19(2):409–426 (1982)

[51] Vavasis, S. A., Zippel, R.: Proving polynomial time for sphere-constrained quadratic programming. Technical report, Department of Computer Science, Cornell University, 90-1182 (1990)

[52] Wang, L., Kim, Y., Li, R.: Calibrating nonconvex penalized regression in ultra-high dimension. *Ann. Stat.*, 5(41):2505–2536 (2013)

[53] Wang, Z., Liu, H., Zhang, T.: Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Stat.*, 6(42):2164–2201 (2014)

[54] Ye, Y.: On affine scaling algorithms for nonconvex quadratic programming. *Mathematical Programming*, 56(1-3):285–300 (1992)

[55] Ye, Y.: On the complexity of approximating a KKT point of quadratic programming. *Mathematical Programming*, 80(2):195–211 (1998)