# Barrier Algorithms for Constrained Non-Convex Optimization

**Pavel Dvurechensky** [1] **Mathias Staudigl** [2]

## Abstract

In this paper, we theoretically show that interior-point methods based on self-concordant barriers possess favorable global complexity beyond their standard application area of convex optimization. To do that we propose first- and second-order methods for non-convex optimization problems with general convex set constraints and linear constraints. Our methods attain a suitably defined class of approximate first- or second-order KKT points with the worst-case iteration complexity similar to unconstrained problems, namely $O(\varepsilon^{-2})$ (first-order) and $O(\varepsilon^{-3/2})$ (second-order), respectively.

## 1. Introduction

Interior-point methods are a universal and very powerful tool for convex optimization (Nesterov & Nemirovskii, 1994; Boyd & Vandenberghe, 2004) that allows one to obtain favorable global complexity guarantees for a variety of problems with many applications. Much less is known about the complexity guarantees for such methods in the non-convex world, especially important in machine learning applications such as training neural networks. This paper aims to fill this gap in the theoretical analysis of interior-point methods in application to optimization with non-convex objectives.

Let $\mathbb{E}$ be a finite-dimensional vector space with inner product $\langle\cdot,\cdot\rangle$ and Euclidean norm $\|\cdot\|$. Our goal is to solve constrained optimization problems of the form

$$\min_x f(x) \quad \text{s.t.: } \mathbf{A}x = b,\ x \in \bar{\mathsf{K}}. \tag{Opt}$$

Our main assumption is as follows:

**Assumption 1.1.**

1. $\bar{\mathsf{K}} \subset \mathbb{E}$ is a closed convex set with nonempty relative interior $\mathsf{K}$;
2. $\mathbf{A} : \mathbb{E} \to \mathbb{R}^m$ is a linear operator assigning each element $x \in \mathbb{E}$ to a vector in $\mathbb{R}^m$ and having full rank, i.e., $\mathrm{im}(\mathbf{A}) = \mathbb{R}^m$, $b \in \mathbb{R}^m$ (all the rows of $\mathbf{A}$ are linearly independent and there are no redundant, linearly dependent constraints);
3. The feasible set $\bar{\mathsf{X}} = \bar{\mathsf{K}} \cap \mathsf{L}$ with $\mathsf{L} = \{x \in \mathbb{E} | \mathbf{A}x = b\}$ has nonempty relative interior denoted by $\mathsf{X} = \mathsf{K} \cap \mathsf{L}$;
4. $f : \mathbb{E} \to \mathbb{R}$ is possibly *non-convex*, continuous on $\bar{\mathsf{X}}$ and continuously differentiable on $\mathsf{X}$;
5. Problem (Opt) admits a global solution. We let $f_{\min}(\mathsf{X}) = \min\{f(x) | x \in \bar{\mathsf{X}}\}$.

As a main tool for developing our algorithms, we use a *self-concordant barrier* (SCB) $h(x)$ for the set $\bar{\mathsf{K}}$, see (Nesterov & Nemirovski, 1994) and Definition 1.2. Using the barrier $h$, our algorithms are designed to reduce the *potential function*

$$F_\mu(x) = f(x) + \mu h(x), \tag{1}$$

where $\mu > 0$ is a (typically) small penalty parameter.

This approach has the following advantages compared to other approaches to solving (Opt).

1. Unlike proximal- or projection-based approaches (Ghadimi & Lan, 2016; Bogolubsky et al., 2016; Cartis et al., 2019; 2012a; Curtis et al., 2017; Cartis et al., 2012b; Dvurechensky, 2017; Birgin & Martínez, 2018; Cartis et al., 2018; 2019), our approach does not require evaluation of a costly projection onto the feasible set $\bar{\mathsf{X}} = \bar{\mathsf{K}} \cap \mathsf{L}$.
2. Unlike splitting methods or augmented Lagrangian algorithms (Birgin & Martínez, 2020; Grapiglia & Yuan, 2020; Andreani et al., 2019; 2021) our algorithms generate feasible approximate solutions and have complexity guarantees similar to the optimal complexity guarantees for unconstraied non-convex optimization.
3. Since our algorithms generate a sequence of relatively interior points, the objective $f$ does not need to be differentiable at the relative boundary. A prominent example of such applications is the nonlinear regression problem with sparsity penalty:

$$\min_{x \geq 0} \left\{ f(x) = \ell(x) + \lambda \|x\|_p^p \right\}, \tag{2}$$

[1] Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany [2] University of Mannheim, Mannheim, Germany. Correspondence to: Pavel Dvurechensky <pavel.dvurechensky@wias-berlin.de>.

*Proceedings of the $41^{st}$ International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

where $\ell(x)$ is a non-convex loss function, $\lambda > 0$, $p \in (0, 1)$. Further motivating applications are discussed in Appendix A.1.

4. Our penalty-based approach provides a flexible framework, since SCBs possess a calculus. Indeed, a sum $h_{\bar{\mathsf{K}}_1} + h_{\bar{\mathsf{K}}_2}$ of SCBs for $\bar{\mathsf{K}}_1$ and $\bar{\mathsf{K}}_2$ is a SCB for $\bar{\mathsf{K}}_1 \cap \bar{\mathsf{K}}_2$. Moreover, SCBs can be efficiently constructed for a large variety of sets encountered in applications (Nesterov, 2018). This is important, for example when $\bar{\mathsf{K}}$ is given as an intersection of 1-norm and Total Variation balls (Liu et al., 2018; Hansen & Bianchi, 2023).

**Related works.** Motivated, in particular, by training of neural networks, non-convex optimization is an active area of research in optimization and ML communities, see, e.g., the review (Danilova et al., 2022) and the references therein. An important part of this research concerns the global complexity guarantees of the proposed algorithms.

**First-order methods.** If $f$ has Lipschitz gradient and there are no constraints, the standard gradient descent achieves the lower iteration complexity bound $O(\varepsilon^{-2})$ to find a first-order $\varepsilon$-stationary point $\hat{x}$ such that $\|\nabla f(\hat{x})\| \leqslant \varepsilon$ (Nesterov, 2018; Carmon et al., 2019b;a). In the composite optimization setting, which includes problems with simple, projection-friendly constraints, a similar iteration complexity is achieved by the mirror descent algorithm (Lan, 2020; Ghadimi et al., 2016; Bogolubsky et al., 2016). Various acceleration strategies of mirror and gradient descent methods have been derived in the literature, attaining the same bound as of gradient descent in the unconstrained case (Ghadimi & Lan, 2016; Guminov et al., 2019; Nesterov et al., 2020; Guminov et al., 2021) or improving upon it under additional assumptions (Carmon et al., 2017; Agarwal et al., 2017). A potential drawback of these algorithms is the computationally expensive projection onto the set $\bar{\mathsf{X}} = \bar{\mathsf{K}} \cap \mathsf{L}$.

**Second-order methods.** If $f$ has Lipschitz Hessian and there are no constraints, cubic-regularized Newton method (Griewank, 1981; Nesterov & Polyak, 2006) and second-order trust region algorithms (Conn et al., 2000; Cartis et al., 2012a; Curtis et al., 2017) achieve the lower iteration complexity bound $O(\max\{\varepsilon_1^{-3/2}, \varepsilon_2^{-3/2}\})$ (Carmon et al., 2019b;a) to find a second-order $(\varepsilon_1, \varepsilon_2)$-stationary point $\hat{x}$ such that $\|\nabla f(\hat{x})\|_2 \leq \varepsilon_1$ and $\lambda_{\min}\left(\nabla^2 f(\hat{x})\right) \geq -\sqrt{\varepsilon_2}$, where $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue of a matrix [1]. Extensions for problems with simple projection-friendly constraints also exist (Cartis et al., 2012b; Birgin & Martínez, 2018; Cartis et al., 2018) with the same iteration complexity bounds, as well as for problems with nonlinear equality and/or inequality constraints (Curtis et al., 2018; Hinder & Ye, 2018; Cartis et al., 2019; Birgin & Martínez, 2020; Grapiglia & Yuan, 2020; Xie & Wright, 2019), but these works do not consider general set constraints as in (Opt) and again may require projections on $\bar{\mathsf{X}} = \bar{\mathsf{K}} \cap \mathsf{L}$.

**Barrier algorithms.** Existing barrier methods for non-convex optimization deal with some particular cases of (Opt), such as $\bar{\mathsf{K}}$ being the non-negative orthant (Ye, 1992; Tseng et al., 2011; Bomze et al., 2019; Bian et al., 2015; Haeser et al., 2019; O'Neill & Wright, 2020), or a symmetric cone (He & Lu, 2022; Dvurechensky & Staudigl, 2024) and $f$ being a quadratic function (Ye, 1992; Faybusovich & Lu, 2006; Lu & Yuan, 2007). None of these works covers the general problem (Opt).

Summarizing, the existing works require at least one of the following assumptions: a) Lipschitz continuity of the gradient and/or Hessian on the whole feasible set, b) no constraints, or simple convex constraints that allow an easy projection, or constraints that do not involve general feasible sets, e.g., only non-negativity constraints. Moreover, existing algorithms do not always come with complexity guarantees. For further discussion of the related literature, see Appendix A.2.

In this paper we develop a flexible and unifying algorithmic framework that is able to accommodate first- and second-order interior-point algorithms for (Opt) with potentially non-convex and non-smooth at the relative boundary objective functions, and general set constraints. To the best of our knowledge, our framework is the first one providing complexity results for first- and second-order algorithms to reach points satisfying, respectively, suitably defined approximate first- and second-order necessary optimality conditions, under such weak assumptions and for such a general setting.

**Contributions.** In this paper, we close the gap in the theoretical analysis of barrier algorithms for general non-convex objectives and general set constraints by constructing first- and second-order algorithms with complexities $O(\varepsilon^{-1})$ and $O(\varepsilon^{-3/2})$ respectively. In more detail, our contributions are as follows:

**Optimality conditions.** We propose a suitable set of first- and second-order necessary optimality conditions for (Opt) that do not require $f$ to be differentiable at the relative boundary of $\bar{\mathsf{K}}$. This is followed by the definition of approximate stationary points which we call $\varepsilon$-KKT and $(\varepsilon_1, \varepsilon_2)$-2KKT points respectively (see Section 2 for a precise definition).

**First-order algorithm.** We propose a new *first-order adaptive barrier method* (**FOABM**, Algorithm 1). To find a

[1] A number of works, e.g. (Cartis et al., 2012a; O'Neill & Wright, 2020), consider an $(\varepsilon_1, \varepsilon_2)$-stationary point defined as $\hat{x}$ such that $\|\nabla f(\hat{x})\|_2 \leq \varepsilon_1$ and $\lambda_{\min}\left(\nabla^2 f(\hat{x})\right) \geq -\varepsilon_2$ and the corresponding complexity $O(\max\{\varepsilon_1^{-3/2}, \varepsilon_2^{-3}\})$. Our definition and complexity bound are the same up to the redefinition of $\varepsilon_2$.

step direction, a linear model for $F_\mu$ in (1), regularized by the squared local norm induced by the Hessian of $h$, is minimized over the tangent space of the linear subspace $\mathsf{L}$. We prove that our new first-order method enjoys (see Theorem 3.3) the upper iteration complexity bound $O(\varepsilon^{-2})$ for reaching an $\varepsilon$-KKT point when a "descent Lemma" holds relative to the local norm induced by the Hessian of $h$ (see Assumption 3.1 for precise definition). Our algorithm is adaptive in the sense that it does not require the knowledge of the Lipschitz constant of the gradient.

**Second-order algorithm.** We propose a new *second-order adaptive barrier method* (**SOABM**, Algorithm 2), for which the step direction is determined by a minimization subproblem over the same tangent space. But, in this case, the minimized model is composed of the linear model for $F_\mu$ augmented by second-order term for $f$ and regularized by the cube of the local norm induced by the Hessian of $h$. The regularization parameter is chosen adaptively in the spirit of (Nesterov & Polyak, 2006; Cartis et al., 2012b). The resulting minimization subproblem can be formulated as a non-convex optimization problem over a linear subspace, which can be solved as in the unconstrained case originally studied in (Nesterov & Polyak, 2006). We establish (see Theorem 4.4) the worst-case bound $O(\max\{\varepsilon_1^{-3/2},\varepsilon_2^{-3/2}\})$ on the number of iterations to reach an $(\varepsilon_1,\varepsilon_2)$-2KKT point under a weaker version of assumption that the Hessian of $f$ is Lipschitz relative to the local norm induced by the Hessian of $h$ (see Assumption 4.1).

We do not perform numerical experiments for two reasons. 1) Our main goal is theoretical and we show that barrier algorithms possess favorable complexity beyond convexity. 2) We are not aware of any baseline algorithms with similar complexity that can accommodate with general set constraints, and produce feasible iterates without involving a projection step.

**Notation.** $\mathbb{E}$ denotes a finite-dimensional real vector space, and $\mathbb{E}^*$ the dual space, which is formed by all linear functions on $\mathbb{E}$. The value of $s\in\mathbb{E}^*$ at $x\in\mathbb{E}$ is denoted by $\langle s,x\rangle$. The gradient and Hessian of a (twice) differentiable function $f:\mathbb{E}\to\mathbb{R}$ at $x\in\mathbb{E}$ are denoted as $\nabla f(x)\in\mathbb{E}^*$ and $\nabla^2 f(x)$ respectively. The directional derivative of function $f:\mathbb{E}\to\mathbb{R}$ is defined in the usual way: $Df(x)[v]=\lim_{\varepsilon\to 0+}\frac{1}{\varepsilon}[f(x+\varepsilon v)-f(x)]$. More generally, for $v_1,\ldots,v_p\in\mathbb{E}$, we define $D^p f(x)[v_1,\ldots,v_p]$ the $p$-th directional derivative at $x$ along directions $v_i\in\mathbb{E}$. In that way we define the gradient $\nabla f(x)\in\mathbb{E}^*$ by $Df(x)[u]=\langle\nabla f(x),u\rangle$ and the Hessian $\nabla^2 f(x):\mathbb{E}\to\mathbb{E}^*$ by $\langle\nabla^2 f(x)u,v\rangle = D^2 f(x)[u,v]$. For an operator $\mathbf{H}:\mathbb{E}\to\mathbb{E}^*$, we denote by $\mathbf{H}^*:\mathbb{E}\to\mathbb{E}^*$ its adjoint operator, defined by the identity $(\forall u,v\in\mathbb{E}):\quad\langle\mathbf{H}u,v\rangle = \langle u,\mathbf{H}^*v\rangle$. An operator $\mathbf{H}:\mathbb{E}\to\mathbb{E}^*$ is positive semi-definite if $\langle\mathbf{H}u,u\rangle\ge 0$ for all $u\in\mathbb{E}$, denoted as $\mathbf{H}\succeq 0$. If the inequality is always strict for non-zero $u$, then $\mathbf{H}$ is called positive definite, and we write $\mathbf{H}\succ 0$. The standard Euclidean norm of a vector $x\in\mathbb{E}$ is denoted by $\|x\|$. We denote by $\mathsf{L}_0\triangleq\{v\in\mathbb{E}|\mathbf{A}v=0\}\triangleq\ker(\mathbf{A})$ the tangent space of the affine subspace $\mathsf{L}=\{x|\mathbf{A}x=b\}$.

**Self-concordant barriers.** By our assumptions $\bar{\mathsf{K}}\subset\mathbb{E}$ has a self-concordant barrier $h(x)$ with finite parameter value $\nu$ (Nesterov & Nemirovski, 1994).

**Definition 1.2.** A function $h:\bar{\mathsf{K}}\to(-\infty,\infty]$ with $\operatorname{dom} h=\mathsf{K}$ is called a *$\nu$-self-concordant barrier* ($\nu$-SCB) for the set $\bar{\mathsf{K}}$ if for all $x\in\mathsf{K}$ and $u\in\mathbb{E}$

$$|D^3h(x)[u,u,u]|\le 2D^2h(x)[u,u]^{3/2},\ \text{and}$$

$$\sup_{u\in\mathbb{E}}|2Dh(x)[u]-D^2h(x)[u,u]|\le\nu.$$

We denote the set of $\nu$-self-concordant barriers by $\mathcal{H}_\nu(\mathsf{K})$.

Note that $\nu\ge 1$ and the Hessian $H(x)\triangleq\nabla^2h(x):\mathbb{E}\to\mathbb{E}^*$ is a positive definite linear operator that defines the *local norm* $\|u\|_x\triangleq\langle H(x)u,u\rangle^{1/2}$. The corresponding *dual local norm* on $\mathbb{E}^*$ is then defined as $\|s\|_x^*\triangleq\langle[H(x)]^{-1}s,s\rangle^{1/2}$.

The *Dikin ellipsoid* with center $x\in\mathsf{K}$ and radius $r>0$ is defined as $\mathcal{W}(x;r)\triangleq\{u\in\mathbb{E}|\ \|u-x\|_x<r\}$. This object allows us to guarantee the feasibility of the iterates in each iteration of our algorithms.

**Lemma 1.3** (Theorem 5.1.5 (Nesterov, 2018))**.** *For all $x\in\mathsf{K}$ we have $\mathcal{W}(x;1)\subseteq\mathsf{K}$.*

The following upper bound for the barrier $h$ is used to establish per-iteration decrease of the potential $F_\mu$.

**Proposition 1.4** (Theorem 5.1.9 (Nesterov, 2018))**.** *Let $h\in\mathcal{H}_\nu(\mathsf{K})$, $x\in\operatorname{dom} h$, and a fixed direction $d\in\mathbb{E}$. For all $t\in[0,\frac{1}{\|d\|_x})$, with the convention that $\frac{1}{\|d\|_x}=+\infty$ if $\|d\|_x=0$, we have:*

$$h(x+td)\le h(x)+t\langle\nabla h(x),d\rangle+t^2\|d\|_x^2\omega(t\|d\|_x),\quad(3)$$

*where $\omega(t)=\frac{-t-\ln(1-t)}{t^2}$.*

We will also use the following inequality for the function $\omega(t)$ (Nesterov, 2018, Lemma 5.1.5):

$$\omega(t)\le\frac{1}{2(1-t)},\ t\in[0,1).\quad(4)$$

Appendix B contains some more technical properties of SCBs which are relevant for the proofs.

## 2. Approximate Optimality Conditions

Following (Burachik et al., 1997), for a given $\varepsilon>0$, we define the $\varepsilon$-approximate normal cone for the set $\bar{\mathsf{K}}$ at $x\in\bar{\mathsf{K}}$

as the set

$$\mathsf{NC}^{\varepsilon}_{\bar{\mathsf{K}}}(x) \triangleq \left\{s \in \mathbb{E}^* | \langle s, y - x\rangle \leq \varepsilon \; \forall \; y \in \bar{\mathsf{K}}\right\}. \quad (5)$$

Clearly, we have $\mathsf{NC}^{0}_{\bar{\mathsf{K}}}(x) = \mathsf{NC}_{\bar{\mathsf{K}}}(x)$, where the latter denotes the normal cone for $\bar{\mathsf{K}}$ at $x$.

The following result gives a necessary optimality condition for (Opt). Importantly, this result holds even if $x^*$ is at the boundary of $\bar{\mathsf{K}}$ and $f$ is not differentiable at $x^*$.

**Theorem 2.1.** *Let the assumptions described above hold for problem* (Opt) *and $x^* \in \bar{\mathsf{K}}$ be a local solution to this problem. Then, there exists a sequence of approximate solutions $x^k \in \mathbb{E}$ and sequences of approximate Lagrange multipliers $y^k \in \mathbb{R}^m$, $s^k \in \mathbb{E}^*$ s.t.:*

1. *$x^k \in \mathsf{K}$, $\mathbf{A}x^k = b$ for all $k$ and $x^k \to x^*$,*
2. *$\nabla f(x^k) - \mathbf{A}^* y^k - s^k \to 0$,*
3. *$-s^k \in \mathsf{NC}^{\sigma_k}_{\bar{\mathsf{K}}}(x^k)$, where $\sigma_k \to 0$.*

*If in addition $f$ is twice differentiable on $\mathsf{K}$, then there exist $\theta_k, \delta_k \in (0, \infty)$ such that $\theta_k, \delta_k \to 0$ and*

$$\langle (\nabla^2 f(x^k) + \theta_k H(x^k) + \delta_k \mathbf{I})d, d\rangle \geq 0 \quad (6)$$

*for all $d \in \mathsf{L}_0$, $\mathbf{I} = \mathbf{I}_{\mathbb{E}}$ being the identity operator.*

The proof is based on interior-penalty arguments sketched below, and fully detailed in Appendix C. Namely, there exists a sequence $x^k \to x^*$ that solves the sequence of penalized problems

$$\min_x f(x) + \frac{1}{4}\|x - x^*\|^4 + \mu_k h(x) \quad \text{s.t.:} \;\; \mathbf{A}x = b, \quad (7)$$

where $\mu_k > 0$, $\mu_k \to 0$. The first- and second-order optimality conditions for the latter problem imply then the statement of the Theorem.

The most interesting part of the above result for us is the second-order condition (6) since it allows us to certify second-order stationary points using the Hessian $H(x)$ of the barrier. At the same time, the first-order conditions may be strengthened compared to the ones in Theorem 2.1. Indeed, if $x^*$ is a local solution of the optimization problem (Opt) at which the objective function $f$ is continuously differentiable, then there exists $y^* \in \mathbb{R}^m$ such that $\nabla f(x^*) - \mathbf{A}^* y^* \in -\mathsf{NC}_{\bar{\mathsf{K}}}(x^*)$, or, equivalently,

$$\langle \nabla f(x^*) - \mathbf{A}^* y^*, x - x^*\rangle \geq 0 \qquad \forall x \in \bar{\mathsf{K}}. \quad (8)$$

The standard way to construct an approximate first-order optimality condition is to add an $\varepsilon$-perturbation in the r.h.s. of (8):

$$\langle \nabla f(\bar{x}) - \mathbf{A}^* \bar{y}, x - \bar{x}\rangle \geq -\varepsilon \qquad \forall x \in \bar{\mathsf{K}}. \quad (9)$$

which is equivalent to $-(\nabla f(\bar{x}) - \mathbf{A}^* \bar{y}) \in \mathsf{NC}^{\varepsilon}_{\bar{\mathsf{K}}}(\bar{x})$.

Motivated by the above, we introduce the following notion of an approximate first-order KKT point for problem (Opt).

**Definition 2.2.** Given $\varepsilon \geq 0$, a point $\bar{x} \in \mathbb{E}$ is an $\varepsilon$-KKT point for problem (Opt) if there exists $\bar{y} \in \mathbb{R}^m$ such that

$$\mathbf{A}\bar{x} = b, \bar{x} \in \mathsf{K}, \quad (10)$$

$$\langle \nabla f(\bar{x}) - \mathbf{A}^* \bar{y}, x - \bar{x}\rangle \geq -\varepsilon \qquad \forall x \in \bar{\mathsf{K}}. \quad (11)$$

We underline that when $\varepsilon_k \to 0$, every convergent subsequence $(x^{k_j})_{j\geq 1}$ of a sequence $(x^k)_{k\geq 1}$ of $\varepsilon_k$-KKT points converges to a stationary point in the sense of Theorem 2.1. Indeed, clearly, such subsequence satisfies item 1. After defining $s^{k_j} = \nabla f(x^{k_j}) - \mathbf{A}^* y^{k_j}$, we see that item 2 trivially holds as equality $\nabla f(x^{k_j}) - \mathbf{A}^* y^{k_j} - s^{k_j} = 0$. Finally, the definition of $s^{k_j}$, (11), (5), and the condition $\varepsilon_k \to 0$ imply item 3 with $\sigma_{k_j} = \varepsilon_{k_j}$. Thus, the limit of the subsequence $x^{k_j}$ satisfies the first three items of Theorem 2.1, and thus is a first-order stationary point according to this theorem.

Based on the second-order condition (6) in Theorem 2.1, we can augment Definition 2.2 with an approximate second-order condition. This leads us to the following notion of an approximate second-order KKT point for problem (Opt).

**Definition 2.3.** Given $\varepsilon_1, \varepsilon_2 \geq 0$, a point $\bar{x} \in \mathbb{E}$ is an $(\varepsilon_1, \varepsilon_2)$-2KKT point for problem (Opt) if there exists $\bar{y} \in \mathbb{R}^m$ such that

$$\mathbf{A}\bar{x} = b, \bar{x} \in \mathsf{K}, \quad (12)$$

$$\langle \nabla f(\bar{x}) - \mathbf{A}^* \bar{y}, x - \bar{x}\rangle \geq -\varepsilon_1 \qquad \forall x \in \bar{\mathsf{K}}, \quad (13)$$

$$\nabla^2 f(\bar{x}) + \sqrt{\varepsilon_2} H(\bar{x}) \succeq 0 \;\text{ on }\; \mathsf{L}_0 = \{v \in \mathbb{E} | \mathbf{A}v = 0\}. \quad (14)$$

Note that our definition of an approximate second-order KKT point is motivated by the notion of weak second-order approximate stationary conditions for non-convex optimization using barrier algorithms (Haeser et al., 2019; O'Neill & Wright, 2020; He & Lu, 2022). Just as for Definition 2.2, we can prove that every accumulation point of a sequence of $(\varepsilon_{1,k}, \varepsilon_{2,k})$-2KKT points satisfies items 1, 2, 3 of Theorem 2.1. Setting $\theta_{k_j} = \sqrt{\varepsilon_{2,k_j}}$ and $\delta_k = 0$, we see that the condition (6) also holds. Thus, the limit of the subsequence $x^{k_j}$ satisfies all the four items of Theorem 2.1, and thus is a second-order stationary point according to this theorem. An important advantage of the above definitions is that $\bar{x}$ lies in the relative interior of the feasible set. Thus, $f$ may be non-differentiable at the relative boundary of the feasible set, see, e.g., problem (2). Moreover, we can use the Hessian of the barrier $H(\bar{x})$ in the second-order condition since $\bar{x}$ is in the interior of $\bar{\mathsf{K}}$.

## 3. First-Order Barrier Algorithm

In this section we introduce our first-order potential reduction method for solving (Opt) that uses a barrier $h \in \mathcal{H}_\nu(\mathsf{K})$ and potential function (1).

### 3.1. Smoothness Assumption

Given $x \in \mathsf{X}$, define the set of *feasible directions* as $\mathcal{F}_x \triangleq \{v \in \mathbb{E} | x + v \in \mathsf{X}\}$. Lemma 1.3 implies that

$$\mathcal{T}_x \triangleq \{v \in \mathbb{E} | \mathbf{A}v = 0, \|v\|_x < 1\} \subseteq \mathcal{F}_x. \tag{15}$$

Upon defining $d = [H(x)]^{1/2}v$ for $v \in \mathcal{T}_x$, we obtain a direction $d$ satisfying $\mathbf{A}[H(x)]^{-1/2}d = 0$ and $\|d\| = \|v\|_x$. Hence, for $x \in \mathsf{K}$, we can equivalently characterize the set $\mathcal{T}_x$ as $\mathcal{T}_x = \{[H(x)]^{-1/2}d | \mathbf{A}[H(x)]^{-1/2}d = 0, \|d\| < 1\}$. For the analysis of the first-order algorithm we use the following first-order smoothness condition.

**Assumption 3.1** (Local smoothness)**.** $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ is continuously differentiable on $\mathsf{X}$ and there exists a constant $M > 0$ such that for all $x \in \mathsf{X}$ and $v \in \mathcal{T}_x$, where $\mathcal{T}_x$ is defined in (15), we have

$$f(x + v) - f(x) - \langle \nabla f(x), v \rangle \leq \frac{M}{2}\|v\|_x^2. \tag{16}$$

*Remark* 3.2. If the set $\bar{\mathsf{X}}$ is bounded, we have $\lambda_{\min}(H(x)) \geq \sigma$ for some $\sigma > 0$. In this case, assuming $f$ has an $M$-Lipschitz continuous gradient, the classical descent lemma (Nesterov, 2018) implies Assumption 3.1. Indeed,

$$f(x + v) - f(x) - \langle \nabla f(x), v \rangle \leq \frac{M}{2}\|v\|^2 \leq \frac{M}{2\sigma}\|v\|_x^2. \quad \diamond$$

Considering $x \in \mathsf{X}, v \in \mathcal{T}_x$ and combining eq. (16) with eq. (3) (with $d = v$ and $t = 1 < \frac{1}{\|v\|_x}$) gives us the following upper bound that holds for all $x \in \mathsf{X}, v \in \mathcal{T}_x$ and $L \geq M$

$$\begin{aligned} F_\mu(x + v) \leq & F_\mu(x) + \langle \nabla F_\mu(x), v \rangle + \frac{L}{2}\|v\|_x^2 \\ & + \mu\|v\|_x^2 \omega(\|v\|_x). \end{aligned} \tag{17}$$

### 3.2. Algorithm and Its Complexity

We assume that our algorithm starts from a $\nu$-analytic center, i.e. a point $x^0 \in \mathsf{X}$ such that

$$h(x) \geq h(x^0) - \nu \qquad \forall x \in \mathsf{X}. \tag{18}$$

Obtaining such a point requires solving a *convex* optimization problem $\min_{x \in \mathsf{X}} h(x)$ up to a very loose accuracy $\nu \geq 1$. We denote $\Delta_0^f \triangleq f(x^0) - f_{\min}(\mathsf{X})$.

**Defining the search direction.** Let $x \in \mathsf{X}$ be given. Our first-order method uses a quadratic model

$$Q_\mu^{(1)}(x, v) \triangleq F_\mu(x) + \langle \nabla F_\mu(x), v \rangle + \frac{1}{2}\|v\|_x^2$$

to compute a search direction $v_\mu(x)$, given by

$$v_\mu(x) \triangleq \underset{v \in \mathbb{E}: \mathbf{A}v = 0}{\operatorname{argmin}} \; Q_\mu^{(1)}(x, v). \tag{19}$$

This search direction is determined by the following system of optimality conditions involving the dual variable $y_\mu(x) \in \mathbb{R}^m$:

$$\nabla F_\mu(x) + H(x)v_\mu(x) - \mathbf{A}^* y_\mu(x) = 0, \tag{20}$$

$$\mathbf{A}v_\mu(x) = 0. \tag{21}$$

Since $H(x) \succ 0$ for $x \in \mathsf{X}$, any standard solution method (Nocedal & Wright, 2000) can be applied for the above linear system. Since $H(x) \succ 0$ for $x \in \mathsf{X}$, and $\mathbf{A}$ has full row rank, the optimality conditions have a unique solution.

**Defining the step-size.** Consider a point $x \in \mathsf{X}$ and a point $x^+(t) \triangleq x + tv_\mu(x)$, where $t \geq 0$ is the step-size. Our aim is to choose $t$ to ensure the feasibility of iterates and decrease of the potential. By Lemma 1.3 and (21), we know that $x^+(t) \in \mathsf{X}$ for all $t \in I_{x,\mu} \triangleq [0, \frac{1}{\|v_\mu(x)\|_x})$. Multiplying (20) by $v_\mu(x)$ and using (21), we obtain $\langle \nabla F_\mu(x), v_\mu(x) \rangle = -\|v_\mu(x)\|_x^2$. Choosing $t \in I_{x,\mu}$, we have

$$t^2\|v_\mu(x)\|_x^2 \omega(t\|v_\mu(x)\|_x) \overset{(4)}{\leq} \frac{t^2\|v_\mu(x)\|_x^2}{2(1 - t\|v_\mu(x)\|_x)}.$$

Therefore, if $t\|v_\mu(x)\|_x \leq 1/2$, we get from (17) that

$$\begin{aligned} & F_\mu(x^+(t)) - F_\mu(x) \\ & \leq -t\|v_\mu(x)\|_x^2 + \frac{t^2 M}{2}\|v_\mu(x)\|_x^2 + \mu t^2\|v_\mu(x)\|_x^2 \\ & = -t\|v_\mu(x)\|_x^2 \left(1 - \frac{M + 2\mu}{2}t\right) \triangleq -\eta_x(t). \end{aligned} \tag{22}$$

The function $\eta_x(t)$ is strictly concave with the unique maximum at $\frac{1}{M+2\mu}$. Thus, maximizing the per-iteration decrease $\eta_x(t)$ under the restriction $0 \leq t \leq \frac{1}{2\|v_\mu(x)\|_x}$, we choose the step-size

$$\mathtt{t}_{\mu,M}(x) \triangleq \min\left\{\frac{1}{M + 2\mu}, \frac{1}{2\|v_\mu(x)\|_x}\right\}.$$

**Adaptivity to the Lipschitz constant.** To get rid of the explicit dependence of the step size on the Lipschitz parameter $M$, we propose a backtracking/adaptive procedure in the spirit of (Nesterov & Polyak, 2006). This procedure generates a sequence of positive numbers $(L_k)_{k \geq 0}$

**Algorithm 1:** First-Order Adaptive Barrier Method - $\mathbf{FOABM}(\mu, \varepsilon, L_0, x^0)$

```
Data: $h \in \mathcal{H}_\nu(\mathsf{K}), \mu > 0, \varepsilon > 0, L_0 > 0, x^0 \in \mathsf{X}$.
Result: $(x^k, y^k, s^k, L_k) \in \mathsf{X} \times \mathbb{R}^m \times \mathbb{E}^* \times \mathbb{R}_+$,
        where $s^k = \nabla f(x^k) - \mathbf{A}^* y^k$, and $L_k$ is
        the last estimate of the Lipschitz constant.
Set $k = 0$;
repeat
    Set $i_k = 0$. Find $v^k \triangleq v_\mu(x^k)$ and the
     corresponding dual variable $y^k \triangleq y_\mu(x^k)$ as
     the solution to
        $$\min_{v \in \mathbb{E}: \mathbf{A}v = 0} \{F_\mu(x^k) + \langle \nabla F_\mu(x^k), v \rangle + \frac{1}{2}\|v\|^2_{x^k}\}. \quad (23)$$
    repeat
        $$\alpha_k \triangleq \min\left\{\frac{1}{2^{i_k} L_k + 2\mu}, \frac{1}{2\|v^k\|_{x^k}}\right\} \quad (24)$$
        Set $z^k = x^k + \alpha_k v^k$, $i_k = i_k + 1$;
    until
$$f(z^k) \leq f(x^k) + \langle \nabla f(x^k), z^k - x^k \rangle + 2^{i_k - 1} L_k \|z^k - x^k\|^2_{x^k}. \quad (25)$$
     ;
    Set $L_{k+1} = 2^{i_k - 1} L_k$, $x^{k+1} = z^k$, $k = k + 1$;
until $\|v^k\|_{x^k} < \frac{\varepsilon}{3\nu}$;
```

for which the local Lipschitz smoothness condition (16) holds. More specifically, let $x^k$ be the current position of the algorithm with the corresponding initial local Lipschitz estimate $L_k$ and $v^k = v_\mu(x^k)$ is the corresponding search direction. To determine the next iterate $x^{k+1}$, we iteratively try step-sizes $\alpha_k$ of the form $\mathtt{t}_{\mu, 2^{i_k} L_k}(x^k)$ for $i_k \geq 0$ until the local smoothness condition (16) holds with $x = x^k$, $v = \alpha_k v^k$ and local Lipschitz estimate $M = 2^{i_k} L_k$, see (25). This process must terminate in finitely many steps since when $2^{i_k} L_k \geq M$, inequality (16) with $M$ changed to $2^{i_k} L_k$, i.e., (25), follows from Assumption 3.1. Setting $L_{k+1} = 2^{i_k - 1} L_k$ allows $L_k$ to adaptively decrease so that in the areas where $f$ is more smooth the algorithm uses larger step-sizes.

Combining the definition of the search direction (19) with the above backtracking strategy, yields a First-Order Adaptive Barrier Method (**FOABM**, Algorithm 1).

**Complexity bound.** Our main result on the iteration complexity of Algorithm 1 is the following Theorem.

**Theorem 3.3.** *Let Assumptions 1.1 and 3.1 hold. Fix the error tolerance $\varepsilon > 0$, the regularization parameter $\mu = \frac{\varepsilon}{\nu}$, and some initial guess $L_0 > 0$ for the Lipschitz constant in (16). Let $(x^k)_{k \geq 0}$ be the trajectory generated by $\mathbf{FOABM}(\mu, \varepsilon, L_0, x^0)$, where $x^0$ is a $\nu$-analytic center satisfying (18). Then the algorithm stops in no more than*

$$\mathbb{K}_I(\varepsilon, x^0) = \left\lceil 36(\Delta_0^f + \varepsilon) \frac{\nu^2(\max\{M, L_0\} + \varepsilon/\nu)}{\varepsilon^2} \right\rceil \quad (26)$$

*outer iterations, and the number of inner iterations is no more than $2(\mathbb{K}_I(\varepsilon, x^0) + 1) + \max\{\log_2(M/L_0), 0\}$. Moreover, the last iterate obtained from $\mathbf{FOABM}(\mu, \varepsilon, L_0, x^0)$ constitutes a $2\varepsilon$-KKT point for problem (Opt) in the sense of Definition 2.2.*

*Remark* 3.4. Since $\nu \geq 1$, $\Delta_0^f$ is expected to be larger than $\varepsilon$, and the constant $M$ is potentially large, we see that the main term in the complexity bound (26) is $O\left(\frac{M\nu^2 \Delta_0^f}{\varepsilon^2}\right) = O\left(\frac{1}{\varepsilon^2}\right)$, i.e., has the same dependence on $\varepsilon$ as the standard complexity bounds (Carmon et al., 2019b;a; Lan, 2020) of first-order methods for non-convex problems under the standard Lipschitz-gradient assumption, which on bounded sets is subsumed by our Assumption 3.1. Further, if the function $f$ is linear, Assumption 3.1 holds with $M = 0$ and we can take $L_0 = 0$. In this case, the complexity bound (26) improves to $O\left(\frac{\nu \Delta_0^f}{\varepsilon}\right)$. ◊

*Remark* 3.5. A potential drawback of Algorithm 1 may be that it requires to fix the parameter $\varepsilon$ before start. This may be easily resolved by a restart procedure with warm starts. Namely, we run Algorithm 1 in epochs numbered by $i \geq 0$. For each restart, we choose $x^0$ as the output of the previous restart and set $\varepsilon = \varepsilon_i = 2^{-i}\varepsilon_0$. Such an algorithm may be run infinitely. To reach any desired accuray $\varepsilon$, it is sufficient to make $I = I(\varepsilon) = O(\log_2(\varepsilon_0/\varepsilon))$ restarts. Then, the total number of inner iterations is of the order $\sum_{i=0}^{I-1} \frac{C}{\varepsilon_i^2} = \sum_{i=0}^{I-1} \frac{C 2^i}{\varepsilon_0^2} = O(\varepsilon^{-2})$, i.e., the complexity is the same. ◊

**Sketch of the proof of Theorem 3.3.** The proof is organized in three steps. First, we prove the correctness of the algorithm, i.e., that it generates a sequence of points in $\mathsf{X}$, and, thus, is indeed an interior-point method. This follows from the construction of the algorithm by an induction argument. Next, we show that the line-search process of finding appropriate $L_k$ in each iteration is finite, and estimate the total number of trials in this process. Then we enter the core of our analysis where we prove that, if the stopping criterion does not hold at iteration $k$, i.e., $\|v^k\|_{x^k} \geq \frac{\varepsilon}{3\nu}$, then the objective $f$ is decreased by a quantity $O(\varepsilon^2)$, which follows from (22). Since the objective is globally lower bounded, we conclude that the method stops in at most $O(\varepsilon^{-2})$ iterations. Finally, we show that when the stopping criterion holds, i.e., $\|v^k\|_{x^k} < \frac{\varepsilon}{3\nu}$, the method has generated an $\varepsilon$-KKT point. On a high level, the result follows from the optimality condition (20) which implies by the stopping condition

$$\nabla f(x^k) - \mathbf{A}^* y^k = -H(x^k)v^k - \mu \nabla h(x^k) = O(\varepsilon + \mu\nu).$$

The full proof is deferred to Appendix D.

## 4. Second-Order Barrier Algorithm

In this section, we present a second-order method that uses also the Hessian of $f$ and the following assumption.

**Assumption 4.1** (Local second-order smoothness)**.** $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ is twice continuously differentiable on $\mathsf{X}$ and there exists a constant $M > 0$ such that, for all $x \in \mathsf{X}$ and $v \in \mathcal{T}_x$, where $\mathcal{T}_x$ is defined in (15), we have

$$\|\nabla f(x+v) - \nabla f(x) - \nabla^2 f(x)v\|_x^* \leq \frac{M}{2}\|v\|_x^2. \quad (27)$$

By integration, it is easy to show that a sufficient condition for (27) is the following local counterpart of the global Lipschitz condition on $\nabla^2 f$ (Nesterov, 2018):

$$\|\nabla^2 f(x+u) - \nabla^2 f(x+v)\|_{\mathrm{op},x} \leq M\|u-v\|_x, \quad (28)$$

where $\|\mathbf{B}\|_{\mathrm{op},x} \triangleq \sup_{u:\|u\|_x \leq 1}\left\{\frac{\|\mathbf{B}u\|_x^*}{\|u\|_x}\right\}$ is the induced operator norm for a linear operator $\mathbf{B} : \mathbb{E} \to \mathbb{E}^*$. Further, again by integration (27) implies

$$f(x+v) - \left[f(x) + \langle \nabla f(x), v\rangle + \frac{1}{2}\langle \nabla^2 f(x)v, v\rangle\right] \leq \frac{M}{6}\|v\|_x^3. \quad (29)$$

*Remark* 4.2. Assumption 4.1 subsumes, if $\bar{\mathsf{X}}$ is bounded, the standard Lipschitz-Hessian setting. If the Hessian of $f$ is $M$-Lipschitz w.r.t. the standard Euclidean norm, we have by (Nesterov, 2018), Lemma 1.2.4, that

$$\|\nabla f(x+v) - \nabla f(x) - \nabla^2 f(x)v\| \leq \frac{M}{2}\|v\|^2.$$

Since $\bar{\mathsf{X}}$ is bounded, one can observe that $\lambda_{\max}([H(x)]^{-1})^{-1} = \lambda_{\min}(H(x)) \geq \sigma$ for some $\sigma > 0$, and (27) holds. Indeed, denoting $g = \nabla f(x+v) - \nabla f(x) - \nabla^2 f(x)v$, we obtain

$$(\|g\|_x^*)^2 \leq \lambda_{\max}([H(x)]^{-1})\|g\|^2 \leq \frac{M^2}{4\lambda_{\min}(H(x))}\|v\|^4 \leq \frac{M^2}{4\sigma^3}\|v\|_x^4. \qquad \diamond$$

Assumption 4.1 also implies, via (29) and (3) (with $d = v$ and $t = 1 < \frac{1}{\|v\|_x}$), the following upper bound for $F_\mu$ that holds for all $x \in \mathsf{X}, v \in \mathcal{T}_x$ and $L \geq M$:

$$F_\mu(x+v) \leq F_\mu(x) + \langle \nabla F_\mu(x), v\rangle + \frac{1}{2}\langle \nabla^2 f(x)v, v\rangle + \frac{L}{6}\|v\|_x^3 + \mu\|v\|_x^2\omega(\|v\|_x). \quad (30)$$

### 4.1. Algorithm and Its Complexity

**Defining the search direction.** Let $x \in \mathsf{X}$ be given. In order to find a search direction, we choose a parameter $L > 0$, construct a cubic-regularized model of the potential $F_\mu$ (1)

$$Q_{\mu,L}^{(2)}(x,v) \triangleq F_\mu(x) + \langle \nabla F_\mu(x), v\rangle + \frac{1}{2}\langle \nabla^2 f(x)v, v\rangle + \frac{L}{6}\|v\|_x^3, \quad (31)$$

and minimize it on the linear subspace $\mathsf{L}_0$:

$$v_{\mu,L}(x) \in \operatorname*{Argmin}_{v\in\mathbb{E}:\mathbf{A}v=0} Q_{\mu,L}^{(2)}(x,v), \quad (32)$$

where by $\mathrm{Argmin}$ we denote the set of global minimizers. The model consists of three parts: linear approximation of $h$, quadratic approximation of $f$, and a cubic regularizer with penalty parameter $L > 0$. Since this model and our algorithm use the second derivative of $f$, we call it a second-order method. Our further derivations rely on the first-order optimality conditions for the problem (32), which say that there exists $y_{\mu,L}(x) \in \mathbb{R}^m$ such that $v_{\mu,L}(x)$ satisfies

$$\nabla F_\mu(x) + \nabla^2 f(x)v_{\mu,L}(x) + \frac{L}{2}\|v_{\mu,L}(x)\|_x H(x)v_{\mu,L}(x) - \mathbf{A}^* y_{\mu,L}(x) = 0, \quad (33)$$

$$-\mathbf{A}v_{\mu,L}(x) = 0. \quad (34)$$

We also use the following extension of (Nesterov & Polyak, 2006), Prop. 1, with the local norm induced by $H(x)$.

**Proposition 4.3.** *For all $x \in \mathsf{X}$ it holds*

$$\nabla^2 f(x) + \frac{L}{2}\|v_{\mu,L}(x)\|_x H(x) \succeq 0 \qquad \text{on } \mathsf{L}_0. \quad (35)$$

**Defining the step-size.** To define the step-size, we act in the same fashion as in Section 3 by considering $x^+(t) \triangleq x + tv_{\mu,L}(x)$, where $t \geq 0$ is a step-size. Using the optimality conditions (33), (34), (35), we estimate (the full derivation is given in Appendix E) the progress parameterized by $t$:

$$\begin{aligned} F_\mu(x^+(t)) - F_\mu(x) &\leq -\frac{Lt^2\|v_{\mu,L}(x)\|_x^3}{12}(3-2t) + \mu t^2\|v_{\mu,L}(x)\|_x^2 \\ &= -\|v_{\mu,L}(x)\|_x^3\frac{Lt^2}{12}\left(3 - 2t - \frac{12\mu}{L\|v_{\mu,L}(x)\|_x}\right) \triangleq -\eta_x(t). \end{aligned} \quad (36)$$

The above inequality holds for all $t \geq 0$ s.t. $t\|v_{\mu,L}(x)\|_x \leq 1/2$. To respect these constraints and guarantee that the decrease is positive, we choose the following step-size rule

$$\mathtt{t}_{\mu,L}(x) \triangleq \min\left\{1, \frac{1}{2\|v_{\mu,L}(x)\|_x}\right\}. \quad (37)$$

Note that $\mathtt{t}_{\mu,L}(x) \leq 1$ and $\mathtt{t}_{\mu,L}(x)\|v_{\mu,L}(x)\|_x \leq 1/2$. Thus, this choice of the step-size is feasible to derive (36).

**Adaptivity to the Lipschitz constant.** Just like Algorithm 1, our second-order method employs a line-search procedure to estimate the Lipschitz constant $M$ in (27), (29) in the spirit of (Nesterov & Polyak, 2006; Cartis et al., 2012b). More specifically, suppose that $x^k \in \mathsf{X}$ is the current position of the algorithm with the corresponding initial local Lipschitz estimate $M_k$. To determine the next iterate $x^{k+1}$, we solve problem (32) with $L = L_k = 2^{i_k} M_k$ starting with $i_k = 0$, find the corresponding search direction $v^k = v_{\mu,L_k}(x^k)$ and the new point $x^{k+1} = x^k + \mathtt{t}_{\mu,L_k}(x^k)v^k$. Then, we check whether the inequalities (27) and (29) hold with $M = L_k$, $x = x^k$, $v = \mathtt{t}_{\mu,L_k}(x^k)v^k$, see (41) and (40). If they hold, we make a step to $x^{k+1}$. Otherwise, we increase $i_k$ by 1 and repeat the procedure. Obviously, when $L_k = 2^{i_k} M_k \geq M$, both inequalities (27) and (29) with $M$ changed to $L_k$, i.e., (41) and (40), are satisfied and the line-search procedure ends. For the next iteration we set $M_{k+1} = \max\{2^{i_k - 1} M_k, \underline{L}\} = \max\{L_k/2, \underline{L}\}$, so that the estimate for the local Lipschitz constant on the one hand can decrease allowing larger step-sizes, and on the other hand is bounded from below.

**Algorithm.** Combining the definition of the search direction in (32) with the just outlined backtracking strategy, yields a Second-Order Adaptive Barrier Method ($\mathbf{SOABM}$, Algorithm 2).

**Complexity bound.** Our main result on the iteration complexity of Algorithm 2 is the following Theorem. The proof follows similar steps as the proof of Theorem 3.3 and is given in Appendix E.2.

**Theorem 4.4.** *Let Assumptions 1.1 and 4.1 hold. Fix the error tolerance $\varepsilon > 0$, the regularization parameter $\mu = \frac{\varepsilon}{4\nu}$, and some initial guess $M_0 > 144\varepsilon$ for the Lipschitz constant in (27). Let $(x^k)_{k\geq 0}$ be the trajectory generated by $\mathbf{SOABM}(\mu, \varepsilon, M_0, x^0)$, where $x^0$ is a $4\nu$-analytic center satisfying (18). Then the algorithm stops in no more than*

$$\mathbb{K}_{II}(\varepsilon, x^0) = \left\lceil \frac{576\nu^{3/2}\sqrt{6\max\{M, M_0\}}(\Delta_0^f + \varepsilon)}{\varepsilon^{3/2}} \right\rceil \tag{42}$$

*outer iterations, and the number of inner iterations is no more than $2(\mathbb{K}_{II}(\varepsilon, x^0) + 1) + 2\max\{\log_2(2M/M_0), 1\}$. Moreover, the output of $\mathbf{SOABM}(\mu, \varepsilon, M_0, x^0)$ constitutes an $(\varepsilon, \frac{\max\{M, M_0\}\varepsilon}{24\nu})$-2KKT point for problem (Opt) in the sense of Definition 2.3.*

Note that Algorithm 2 can be made anytime convergent by the same restarting procedure explained in Remark 3.5.

**Algorithm 2:** Second-Order Adaptive Barrier Method - $\mathbf{SOABM}(\mu, \varepsilon, M_0, x^0)$

```
Data: $h \in \mathcal{H}_\nu(\mathsf{K})$,
      $\mu > 0, \varepsilon > 0, M_0 \geq 144\varepsilon, x^0 \in \mathsf{X}$.
Result: $(x^k, y^{k-1}, s^k, M_k) \in \mathsf{X} \times \mathbb{R}^m \times \mathsf{K}^* \times \mathbb{R}_+$,
        where $s^k = \nabla f(x^k) - \mathbf{A}^* y^{k-1}$, and $M_k$ is
        the last estimate of the Lipschitz constant.
Set $\underline{L} \triangleq 144\varepsilon$, $k = 0$;
repeat
    Set $i_k = 0$.
    repeat
        Set $L_k = 2^{i_k} M_k$. Find $v^k \triangleq v_{\mu,L_k}(x^k)$ and
         $y^k \triangleq y_{\mu,L_k}(x^k)$ as a global solution to
        $$\min_{v:\mathbf{A}v=0} Q^{(2)}_{\mu,L_k}(x^k, v), \quad \text{where } Q^{(2)}_{\mu,L}(x, v) \text{ as in (31).} \tag{38}$$
        Set $\alpha_k \triangleq \min\left\{1, 1/(2\|v^k\|_{x^k})\right\}$.      (39)
        Set $z^k = x^k + \alpha_k v^k$, $i_k = i_k + 1$;
    until
    $$f(z^k) \leq f(x^k) + \langle \nabla f(x^k), z^k - x^k \rangle + \frac{1}{2}\langle \nabla^2 f(x^k)(z^k - x^k), z^k - x^k \rangle + \frac{L_k}{6}\|z^k - x^k\|^3_{x^k}, \tag{40}$$
    and $\|\nabla f(z^k) - \nabla f(x^k) - \nabla^2 f(x^k)(z^k - x^k)\|^*_{x^k} \leq \frac{L_k}{2}\|z^k - x^k\|^2_{x^k}$.      (41)
    ;
    Set $M_{k+1} = \max\{\frac{L_k}{2}, \underline{L}\}$, $x^{k+1} = z^k$,
     $k = k + 1$
until $\|v^{k-1}\|_{x^{k-1}} < \Delta_{k-1} \triangleq \sqrt{\frac{\varepsilon}{12L_{k-1}\nu}}$ and
 $\|v^k\|_{x^k} < \Delta_k \triangleq \sqrt{\frac{\varepsilon}{12L_k\nu}}$;
```

*Remark* 4.5. Since $\Delta_0^f$ is expected to be larger than $\varepsilon$, and the constant $M$ is potentially large, we see that the main term in the complexity bound (42) is $O\left(\frac{\nu^{3/2}\sqrt{M}\Delta_0^f}{\varepsilon^{3/2}}\right) = O(\varepsilon^{-3/2})$. Note that the complexity result $O(\max\{\varepsilon_1^{-3/2}, \varepsilon_2^{-3/2}\})$ reported in (Carmon et al., 2019b;a) to find an $(\varepsilon_1, \varepsilon_2)$-2KKT point for arbitrary $\varepsilon_1, \varepsilon_2 > 0$, is known to be optimal for unconstrained smooth non-convex optimization by second-order methods under the standard Lipschitz-Hessian assumption, subsumed on bounded sets by our Assumption 4.1. A similar dependence on arbitrary $\varepsilon_1, \varepsilon_2 > 0$ can be easily obtained from our theorem by setting $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$. ◊

*Remark* 4.6. An interesting observation is that our algorithm can be interpreted as a damped version of a cubic-regularized Newton's method. We have that the stepsize $\alpha_k$ satisfies $\alpha_k = \min\left\{1, \frac{1}{2\|v^k\|_{x^k}}\right\}$. At the initial phase, when the algorithm is far from an $(\varepsilon_1, \varepsilon_2)$-2KKT point, we have $\|v^k\|_{x^k} > 1/2$ and $\alpha_k = \frac{1}{2\|v^k\|_{x^k}} < 1$. When the algorithm is getting closer to $(\varepsilon_1, \varepsilon_2)$-2KKT point, $\|v^k\|_{x^k}$ becomes smaller and the algorithm automatically switches to full steps $\alpha_k = 1$.

At the same time our algorithm is completely different from cubic-regularized Newton's method (Nesterov & Polyak, 2006) applied to minimize the potential $F_\mu$. Indeed, we regularize by the cube of the local norm, rather than the cube of the standard Euclidean norm, and we do not form a second-order Taylor expansion of $F_\mu$. These adjustments are needed to align the search direction subproblem with the local geometry of the feasible set. Moreover, for our algorithm, the analysis of the cubic-regularized Newton's method is not directly applicable since it relies on stepsize 1, which may lead to infeasible iterates in our case. $\Diamond$

## 5. Conclusion

In this paper, we propose first- and second-order algorithms for non-convex problems with linear and general set constraints. We develop also necessary optimality conditions for such problems and define their suitable approximate counterparts. Further, we show that our algorithms achieve approximate stationary points with ”optimal” worst-case iteration complexity. Unlike previously known results on interior-point methods for non-convex optimization, our approach allows one to solve a much wider class of problems.

So far the interior point methods in the sense of the book (Nesterov & Nemirovskii, 1994) are classical and powerful for convex setting where they are universal, there are a lot of standard solvers based on these methods, and there are human-language solvers like CVX (Boyd & Vandenberghe, 2004), etc. For non-convex setting the study of these ideas are on a case by case basis with many works for many particular cases. We extend the universality property to non-convex setting in the most generality known so far.

Future works include extensions of our algorithms for the setting of inexact solution of search direction finding problems. With that respect we believe that it is possible to construct a Newton-conjugate-gradient counterpart of our second-order method. Further, we plan to use the proposed methods in machine learning applications such as constrained non-linear regression and training Input Convex Neural Networks (Amos et al., 2017). Further potential extensions include adding non-linear functional constraints to the problem.

## Acknowledgements

M. Staudigl acknowledges support from the COST Action CA16228 ”European Network for Game Theory”. The work by P. Dvurechensky was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH$^+$ (EXC-2046/1, project ID: 390685689).

We would like to thank Yurii Nesterov and Geovani Grapiglia for discussions and anonymous reviewers for their fruitful comments.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. Since the work is theoretical and aims at developing new optimization algorithms, we don't see any immediate or potential societal consequences of our work.

## References

Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., and Ma, T. Finding approximate local minima faster than gradient descent. pp. 1195–1199. ACM, 2017. ISBN 145034528X.

Amos, B., Xu, L., and Kolter, J. Z. Input convex neural networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 146–155. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/amos17b.html`.

Andreani, R., Fukuda, E. H., Haeser, G., Santos, D. O., and Secchin, L. D. Optimality conditions for nonlinear second-order cone programming and symmetric cone programming. *Optimization online*, 2019.

Andreani, R., Gómez, W., Haeser, G., Mito, L. M., and Ramos, A. On optimality conditions for nonlinear conic programming. *Mathematics of Operations Research*, 47 (3):2160–2185, 2023/05/13 2021. doi: 10.1287/moor.2021.1203. URL `https://doi.org/10.1287/moor.2021.1203`.

Bach, F. Self-concordant analysis for logistic regression. *Electron. J. Statist.*, 4:384–414, 2010. doi: 10.1214/09-EJS521. URL `https://projecteuclid.org:443/euclid.ejs/1271941980`.

Bian, W., Chen, X., and Ye, Y. Complexity analysis of interior point algorithms for non-lipschitz and nonconvex minimization. *Mathematical Programming*, 149(1):301–327, 2015. doi: 10.1007/s10107-014-0753-5. URL `https://doi.org/10.1007/s10107-014-0753-5`.

Birgin, E. G. and Martínez, J. M. On regularization and active-set methods with complexity for constrained optimization. *SIAM Journal on Optimization*, 28(2): 1367–1395, 2018. doi: 10.1137/17M1127107. URL https://doi.org/10.1137/17M1127107.

Birgin, E. G. and Martínez, J. M. Complexity and performance of an augmented lagrangian algorithm. *Optimization Methods and Software*, 35(5):885–920, 2020. doi: 10.1080/10556788.2020.1746962. URL https://doi.org/10.1080/10556788.2020.1746962.

Bogolubsky, L., Dvurechensky, P., Gasnikov, A., Gusev, G., Nesterov, Y., Raigorodskii, A. M., Tikhonov, A., and Zhukovskii, M. Learning supervised pagerank with gradient-based and gradient-free optimization methods. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 4914–4922. Curran Associates, Inc., 2016. arXiv:1603.00717.

Bomze, I. M., Mertikopoulos, P., Schachinger, W., and Staudigl, M. Hessian barrier algorithms for linearly constrained optimization problems. *SIAM Journal on Optimization*, 29(3):2100–2127, 2019.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, Cambridge, 2004. ISBN 9780521833783. doi: DOI:10.1017/CBO9780511804441. URL https://www.cambridge.org/core/books/convex-optimization/17D2FAA54F641A2F62C7CCD01DFA97C4.

Burachik, R. S., Iusem, A. N., and Svaiter, B. F. Enlargement of monotone operators with applications to variational inequalities. *Set-Valued Analysis*, 5:159–180, 1997.

Carderera, A., Besançon, M., and Pokutta, S. Simple steps are all you need: Frank-wolfe and generalized self-concordant functions. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5390–5401. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/2b323d6eb28422cef49b266557dd31ad-Paper.pdf.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Convex until proven guilty: Dimension-free acceleration of gradient descent on non-convex functions. pp. 654–663. JMLR. org, 2017.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points i. *Mathematical Programming*, 2019a. doi: 10.1007/s10107-019-01406-y. URL https://doi.org/10.1007/s10107-019-01406-y.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 2019b. doi: 10.1007/s10107-019-01431-x. URL https://doi.org/10.1007/s10107-019-01431-x.

Cartis, C., Gould, N. I. M., and Toint, P. L. Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011. doi: 10.1007/s10107-009-0286-5. URL https://doi.org/10.1007/s10107-009-0286-5.

Cartis, C., Gould, N., and Toint, P. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, 28(1):93–108, 2012a. ISSN 0885-064X. doi: https://doi.org/10.1016/j.jco.2011.06.001. URL https://www.sciencedirect.com/science/article/pii/S0885064X11000537.

Cartis, C., Gould, N. I., and Toint, P. L. An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity. *IMA Journal of Numerical Analysis*, 32(4):1662–1695, 2012b.

Cartis, C., Gould, N. I. M., and Toint, P. L. Second-order optimality and beyond: Characterization and evaluation complexity in convexly constrained nonlinear optimization. *Foundations of Computational Mathematics*, 18(5):1073–1107, 2018. doi: 10.1007/s10208-017-9363-y. URL https://doi.org/10.1007/s10208-017-9363-y.

Cartis, C., Gould, N. I. M., and Toint, P. L. Optimality of orders one to three and beyond: characterization and evaluation complexity in constrained nonconvex optimization. *Journal of Complexity*, 53:68–94, 2019.

Conn, A., Gould, N., and Toint, P. *Trust Region Methods*. Society for Industrial and Applied Mathematics, 2000.

Curtis, F. E., Robinson, D. P., and Samadi, M. A trust region algorithm with a worst-case iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming*, 162(1-2):1–32, 2017.

Curtis, F. E., Robinson, D. P., and Samadi, M. Complexity analysis of a trust funnel algorithm for equality constrained optimization. *SIAM Journal on Optimization*, 28 (2):1533–1563, 2018. doi: 10.1137/16M1108650. URL https://doi.org/10.1137/16M1108650.

Danilova, M., Dvurechensky, P., Gasnikov, A., Gorbunov, E., Guminov, S., Kamzolov, D., and Shibaev, I. *Recent Theoretical Advances in Non-Convex Optimization*, pp. 79–163. Springer International Publishing, Cham, 2022. ISBN 978-3-031-00832-0. doi: 10.1007/978-3-031-00832-0_3. URL https://doi.org/10.1007/978-3-031-00832-0_3.

Dvurechensky, P. Gradient method with inexact oracle for composite non-convex optimization. *arXiv:1703.09180*, 2017.

Dvurechensky, P. and Staudigl, M. Hessian barrier algorithms for non-convex conic optimization. *Mathematical Programming*, 2024. doi: 10.1007/s10107-024-02062-7. URL https://doi.org/10.1007/s10107-024-02062-7. arXiv:2111.00100.

Dvurechensky, P., Ostroukhov, P., Safin, K., Shtern, S., and Staudigl, M. Self-concordant analysis of Frank-Wolfe algorithms. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2814–2824, Virtual, 13–18 Jul 2020. PMLR. URL http://proceedings.mlr.press/v119/dvurechensky20a.html. arXiv:2002.04320.

Dvurechensky, P., Safin, K., Shtern, S., and Staudigl, M. Generalized self-concordant analysis of Frank–Wolfe algorithms. *Mathematical Programming*, 198: 255–323, 2023. ISSN 1436-4646. doi: 10.1007/s10107-022-01771-1. URL https://doi.org/10.1007/s10107-022-01771-1. arXiv:2010.01009.

Faybusovich, L. and Lu, Y. Jordan-algebraic aspects of nonconvex optimization over symmetric cones. *Applied Mathematics and Optimization*, 53(1):67–77, 2006. ISSN 1432-0606. URL https://doi.org/10.1007/s00245-005-0835-0.

Fiacco, A. V. and McCormick, G. P. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. John Wiley & Sons, New York, NY, USA, 1968. Reprinted by SIAM Publications in 1990.

Ge, D., Wang, H., Xiong, Z., and Ye, Y. Interior-point methods strike back: Solving the wasserstein barycenter problem. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf.

Ghadimi, S. and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016. doi: 10.1007/s10107-015-0871-8. URL https://doi.org/10.1007/s10107-015-0871-8.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1):267–305, 2016. ISSN 1436-4646. doi: 10.1007/s10107-014-0846-1. URL http://dx.doi.org/10.1007/s10107-014-0846-1.

Grapiglia, G. N. and Yuan, Y.-x. On the complexity of an augmented Lagrangian method for nonconvex optimization. *IMA Journal of Numerical Analysis*, 41 (2):1546–1568, 07 2020. ISSN 0272-4979. doi: 10.1093/imanum/draa021. URL https://doi.org/10.1093/imanum/draa021.

Griewank, A. The modification of newton's method for unconstrained optimization by bounding cubic terms. Technical report, Department of Applied Mathematics and Theoretical Physics, University of Cambridge., 1981.

Guminov, S., Dvurechensky, P., Tupitsa, N., and Gasnikov, A. On a combination of alternating minimization and Nesterov's momentum. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pp. 3886–3898. PMLR, 2021. URL http://proceedings.mlr.press/v139/guminov21a.html.

Guminov, S. V., Nesterov, Y. E., Dvurechensky, P. E., and Gasnikov, A. V. Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems. *Doklady Mathematics*, 99(2): 125–128, Mar 2019.

Haeser, G., Liu, H., and Ye, Y. Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. *Mathematical Programming*, 178(1):263–299, Nov 2019. ISSN 1436-4646. doi: 10.1007/s10107-018-1290-4. URL https://doi.org/10.1007/s10107-018-1290-4.

Hansen, J. B. and Bianchi, F. M. Total variation graph neural networks. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12445–12468. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/hansen23a.html.

Harmany, Z. T., Marcia, R. F., and Willett, R. M. This is spiral-tap: Sparse poisson intensity reconstruction algorithms—theory and practice. *IEEE Transactions on Image Processing*, 21(3):1084–1096, 2011.

He, C. and Lu, Z. A Newton-CG based barrier method for finding a second-order stationary point of nonconvex conic optimization with complexity guarantees. *Forthcoming: SIAM Journal on Optimization*, arXiv:2207.05697, 2022.

Hinder, O. and Ye, Y. Worst-case iteration bounds for log barrier methods for problems with nonconvex constraints. *arXiv:1807.00404*, 2018.

Hong, I., Na, S., Mahoney, M. W., and Kolar, M. Constrained optimization via exact augmented lagrangian and randomized iterative sketching. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13174–13198. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/hong23b.html.

Jia, X., Liang, X., Shen, C., and Zhang, L.-H. Solving the cubic regularization model by a nested restarting lanczos method. *SIAM Journal on Matrix Analysis and Applications*, 43(2):812–839, 2022. doi: 10.1137/21M1436324. URL https://doi.org/10.1137/21M1436324.

Lan, G. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Nature, 2020.

Liu, J., Sun, Y., Xu, X., and Kamilov, U. S. Image restoration using total variation regularized deep image prior. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7715–7719, 2018. URL https://api.semanticscholar.org/CorpusID:53115841.

Loh, P.-L. and Wainwright, M. J. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

Lu, Y. and Yuan, Y. An interior-point trust-region algorithm for general symmetric cone programming. *SIAM Journal on Optimization*, 18(1):65–86, 2020/08/03 2007. doi: 10.1137/040611756. URL https://doi.org/10.1137/040611756.

Marteau-Ferey, U., Ostrovskii, D., Bach, F., and Rudi, A. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. *arXiv preprint arXiv:1902.03046*, 2019.

Monteiro, R., Sicre, M., and Svaiter, B. A hybrid proximal extragradient self-concordant primal barrier method for monotone variational inequalities. *SIAM Journal on Optimization*, 25(4):1965–1996, 2019/09/09 2015. doi: 10.1137/130931862. URL https://doi.org/10.1137/130931862.

Nesterov, Y. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer International Publishing, 2018.

Nesterov, Y. and Nemirovski, A. *Interior Point Polynomial methods in Convex programming*. SIAM Publications, 1994.

Nesterov, Y. and Nemirovskii, A. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 2016/08/20 1994. ISBN 978-0-89871-319-0. doi: doi:10.1137/1.9781611970791. URL http://dx.doi.org/10.1137/1.9781611970791.

Nesterov, Y. and Polyak, B. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. ISSN 1436-4646. doi: 10.1007/s10107-006-0706-8. URL http://dx.doi.org/10.1007/s10107-006-0706-8.

Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, pp. 1–28, 2020. doi: 10.1080/10556788.2020.1731747.

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, 2nd edition, 2000.

Nouiehed, M., Lee, J. D., and Razaviyayn, M. Convergence to second-order stationarity for constrained non-convex optimization. *arXiv preprint arXiv:1810.02024*, 2018.

O'Neill, M. and Wright, S. J. A log-barrier Newton-CG method for bound constrained optimization with complexity guarantees. *IMA Journal of Numerical Analysis*, 12/27/2020 2020. doi: 10.1093/imanum/drz074. URL https://doi.org/10.1093/imanum/drz074.

Tran-Dinh, Q., Sun, T., and Lu, S. Self-concordant inclusions: a unified framework for path-following generalized newton-type algorithms. *Mathematical Programming*, 177(1):173–223, 2019. doi: 10.1007/s10107-018-1264-6. URL https://doi.org/10.1007/s10107-018-1264-6.

Tseng, P., Bomze, I. M., and Schachinger, W. A first-order interior-point method for linearly constrained smooth optimization. *Mathematical Programming*, 127 (2):399–424, 2011. ISSN 1436-4646. doi: 10.1007/s10107-009-0292-7. URL http://dx.doi.org/10.1007/s10107-009-0292-7.

Xie, Y. and Wright, S. J. Complexity of proximal augmented lagrangian for nonconvex optimization with nonlinear equality constraints. *arXiv:1908.00131*, 2019.

Ye, Y. On affine scaling algorithms for nonconvex quadratic programming. *Mathematical Programming*, 56(1):285–300, 1992. doi: 10.1007/BF01580903. URL https://doi.org/10.1007/BF01580903.

Zhang, Y. and Lin, X. Disco: Distributed optimization for self-concordant empirical loss. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 362–370. PMLR, 06 2015. URL http://proceedings.mlr.press/v37/zhangb15.html.

# A. Discussion

In this section, we provide additional details to motivate our work and give concluding remarks. The following sections contain technical details of the proofs.

## A.1. Motivating Applications

Consider a linear inverse problem with forward operator $\Phi$. Our aim is to learn a parameter $u \in U = \mathbb{R}^{n_u}_{+}$ so that the approximate equality

$$\Phi u \approx z$$

holds true. The matrix $\Phi$ has rows $\Phi_1, \ldots, \Phi_d$ and we assume that $\Phi_i \in \mathbb{R}^{n_u}_{+}$. Furthermore, $z \in \mathbb{R}^m_+$ and $\sum_{j=1}^{p} \Phi_{ij} = \phi_j > 0$ for all $j = 1, \ldots, m$. In Poisson linear inverse problems (Harmany et al., 2011), under additional structural assumptions, the penalized maximum likelihood approach for recovering the parameter $u$ leads to minimization of the function

$$\sum_{i=1}^{m} \{(\Phi u)_i - z_i \log((\Phi u)_i)\} + \alpha r(u)$$

where $r(u)$ is a, potentially non-convex, regularizer. An important concrete example is the non-convex sparsity-inducing $\ell_p$ regularizer $r(u) = \sum_i |u_i|^p = \|u\|_p^p$, with $p \in (0, 1)$, an example frequently used in computational statistics (Loh & Wainwright, 2017). We assume that $r$ is continuously differentiable on $\mathrm{int}(U)$ and note that $\|u\|_p^p$ is non-differentiable at $u = 0$. This problem can be written in our optimization template by performing the variable substitution $v = \Phi u$ and imposing the linear constraint

$$\Phi u - v = [\Phi; -I] \begin{pmatrix} u \\ v \end{pmatrix} = 0.$$

Therefore, defining the linear operator $\mathbf{A} = [\Phi; -I]$, we obtain a matrix of rank $m$. We set $x = (u, v)$ and

$$f(x) = \sum_{i=1}^{m} \{v_i - z_i \log(v_i)\} + \alpha r(u)$$

so that our inverse problem of Poisson image recovery reads as

$$\min_{x=(u,v)} f(x) \text{ s.t. } \mathbf{A}x = 0, x \in \bar{\mathsf{K}} = \mathbb{R}^n_+,$$

where $n = n_u + m$. This problem admits the efficient self-concordant barrier

$$h(x) = -\sum_i \log(x_i) \qquad \forall x \in \mathsf{K} = \mathbb{R}^n_{++}.$$

By assumption, the function $f$ is twice continuously differentiable on $\mathbb{R}^n_{++}$. We compute the gradient and the Hessian as

$$\nabla f(x) = \begin{pmatrix} \alpha \nabla r(u) \\ \mathbf{1}_m - V^{-1} z \end{pmatrix}, \text{ and } \nabla^2 f(x) = \begin{pmatrix} \alpha \nabla^2 r(u) & 0 \\ 0 & V^{-2} z \end{pmatrix},$$

where $V = \mathrm{diag}\{v_1, \ldots, v_n\}$. The Hessian matrix of the barrier function decomposes as

$$H(x) = \nabla^2 h(x) = \begin{pmatrix} U^{-2} & 0 \\ 0 & V^{-2} \end{pmatrix}.$$

Thanks to the block structure of the involved matrices, the subproblems involved in the search direction finding routines of our two algorithms can be efficiently handled with efficient numerical linear algebra solvers.

In a similar fashion, we can consider non-linear inverse problems where the loss is given by a squared misfit between the data $z$ and non-linear prediction function $\Phi(x)$ given, e.g., by a neural network. In this case, we have

$$f(x) = \|\Phi(x) - z\|^2 + \alpha r(x).$$

The non-negativity constraints in this case may be motivated by the training of Input Convex Neural Networks (Amos et al., 2017). A related problem of sparse non-linear regression may be reformulated as

$$\min_{x\in\mathbb{R}^n} \|\Phi(x) - z\|^2 \text{ s.t. } \|x\|_1 \leq \lambda.$$

This problem clearly fits our problem template with $f(x) = \|\Phi(x) - z\|^2$, $\mathsf{K} = \{x \in \mathbb{R}^n : \|x\|_1 \leq \lambda\}$ and trivial linear constraint satisfied by all the points in the space $\mathbb{E}$ where $\mathbf{A}$ is a zero row and $b$ is zero.

We now argue why the local Lipschitz continuity assumptions are not unlikely to hold automatically in the above and other machine learning applications. First, loss functions in machine learning are usually coercive, which leads to an implicit compactness of the feasible set. Second, the coercivity of $f$ and the existence of a global solution imply that a solution to the minimization problem is finite and lies in the interior of some ball of some (possibly large) radius. Adding this ball to the set of constraints does not change the solution to the problem and simultaneously leads to the setting of Remark 3.2 since the feasible set is made compact. Third, the coercivity of $f$ implies that the potential function $F_\mu$ is also coercive. Hence, it has bounded (and, hence, compact) level sets. From (51), (69) we see that our algorithms are monotone w.r.t. the potential $F_\mu$, i.e., the potential is decreasing during the optimization process. Hence, algorithms stay on the level set of the potential defined by the starting point. Since this level set is compact and the smoothness assumption is essentially needed on the trajectory of the algorithm which stays on this compact level set, we conclude that Remarks 3.2, 4.2 hold.

### A.2. Additional Comments on Related Literature

Interior-point methods (Ge et al., 2019) and optimization involving self-concordant functions (Bach, 2010; Zhang & Lin, 2015; Tran-Dinh et al., 2019; Marteau-Ferey et al., 2019; Dvurechensky et al., 2020; 2023; Carderera et al., 2021) remains an active area of research in the Machine Learning community. Yet, the main focus of this research stays on solving convex problems. However, the optimization community was recently quite successful in extending interior-point methods from the classical convex world (Nesterov & Nemirovskii, 1994) to non-convex world (Ye, 1992; Tseng et al., 2011; Bomze et al., 2019; Tseng et al., 2011; Bian et al., 2015; Haeser et al., 2019; O'Neill & Wright, 2020; He & Lu, 2022; Dvurechensky & Staudigl, 2024; Ye, 1992; Faybusovich & Lu, 2006; Lu & Yuan, 2007). The benefit of interior-point methods is that when the feasible set is given as an intersection of several sets, these methods allow decomposing the feasible set into separate building blocks. This allows one to avoid expensive projections onto the intersection. Further, such methods guarantee the feasibility of the iterates. At the same time, constrained optimization in the spirit of (Curtis et al., 2018; Hinder & Ye, 2018; Cartis et al., 2019; Birgin & Martínez, 2020; Grapiglia & Yuan, 2020; Xie & Wright, 2019) has recently attracted attention of ML community (Hong et al., 2023) motivated by constrained deep neural networks, physical informed neural networks, PDE-constrained optimization, optimal control, and constrained model estimations.

With this paper, we further narrow the gap between the advances of optimization methods for non-convex problems with complicated constraints and Machine Learning applications. Moreover, our algorithms apply to more general problems than the ones available in the literature on interior-point methods for non-convex optimization, which we describe next and which influenced our work. In a sense, we generalize in this paper this line of works to a much more general class of problems. The authors of (Haeser et al., 2019) propose first- and second-order algorithms with "optimal" [2] complexity guarantees for problems with linear equality constraints and non-negativity constraints, i.e., $\bar{\mathsf{K}} = \mathbb{R}^n_+$. Their algorithms are based on the Trust Region idea, unlike our algorithms that use quadratic and cubic regularization. The authors of (O'Neill & Wright, 2020) consider a similar problem, but without equality constraints. They develop a Newton-conjugate-gradient method with "optimal" complexity to reach a second-order approximate stationary point. The authors of (He & Lu, 2022; Dvurechensky & Staudigl, 2024) consider a more general setting where $\bar{\mathsf{K}}$ is a symmetric cone. (Dvurechensky & Staudigl, 2024) propose first- and second-order methods with "optimal" complexity guarantees, and (He & Lu, 2022) propose a Newton-conjugate-gradient method with "optimal" complexity to reach a second-order approximate stationary point. Importantly, all these papers allow the objective to be non-differentiable at the relative boundary of the feasible set, unlike methods that use projections.

Compared to the optimization template (Opt), all these papers consider a narrower class of problems with $\bar{\mathsf{K}}$ being a cone. Moreover, all the results in these papers heavily rely on the conic structure of the constraints (non-negativity constraints or

[2] Here and below we refer to the complexity bound $O(\varepsilon^{-2})$ for first-order and $O(\max\{\varepsilon_1^{-3/2}, \varepsilon_2^{-3/2}\})$ for second-order methods as "optimal" for two reasons. First, the respective optimal algorithms for unconstrained problems have a similar dependence on the accuracy in the complexity bounds. Second, we are not aware of lower complexity bounds for problems with constraints that we consider. But, it is natural to expect that such problems are no easier than unconstrained problems.

general conic constraints). In particular, the conic structure allows one to easily introduce approximate optimality conditions since conic duality can be used. Further, they use a narrow subclass of logarithmic or more generally logarithmically homogeneous self-concordant barriers that satisfy additionally

$$h(tx) = h(x) - \nu \ln(t) \qquad \forall x \in \operatorname{int}(\mathsf{K}), t > 0,$$

and that are specific for cones and possess additional properties that can be used to derive optimality conditions and complexity results for algorithms. Specifically, (He & Lu, 2022) rely on the structural properties of logarithmically homogeneous barriers to derive optimality conditions and complexity results for their algorithm. The conic feasible set allows them to formulate optimality conditions and their approximate counterparts in terms of the dual cone, which is impossible in our setting. Their optimality conditions are based also on the notion of limiting inverse of the Hessian of the barrier function, an object whose existence again heavily relies on the logarithmic homogeneity of the barrier. This construction is impossible in our setting since in our setting the feasible set is not conic and is not assumed to admit a logarithmically homogeneous barrier. Finally, the proof that their algorithm returns an approximate stationary point also relies on the logarithmic homogeneity property of the barrier. To sum up, each element of their approach is not applicable to our setting since the feasible set is not a cone and thus does not have a logarithmically homogeneous barrier. Our assumptions on the barrier are weaker, since we consider just self-concordant barriers without the additional logarithmic homogeneity property (which essentially forces the domain to be a pointed cone). This simultaneously allows us to solve much more general problems where $\bar{\mathsf{K}}$ is a closed convex set, but not necessarily a cone. The latter in particular is justified by the existence of universal barriers for convex sets (Nesterov & Nemirovski, 1994). We summarize the comparison with related literature in Table A.2.

During the rebuttal phase an anonymous reviewer pointed us to the preprint (Nouiehed et al., 2018) where a general NP-hardness result is proven for checking whether a given point is a $(\varepsilon_1, \varepsilon_2)$-second order stationary point. They establish this result by using a specific criticality measure which explicitly requires bounding the curvature of the objective function in directions $d$ so that the affine translate $x + d$ remains feasible. We (and other related works) claim that the proposed algorithm produces a point that is approximately stationary. Thus, no checking approximate stationarity is involved in our problem. Moreover, our point is not arbitrary since it is obtained by a concrete algorithm. Furthermore, our definition of an approximate second-order stationary point is a *weak* second-order condition which only measures curvature relative to the null space $\mathsf{L}_0$ and the operator $H(\bar{x})$. Hence, our second-order stationary points have a different nature than the ones involved in the NP-hardness result of (Nouiehed et al., 2018). We also draw the reader's attention that the previous literature on barrier methods for non-convex problems (Haeser et al., 2019; O'Neill & Wright, 2020; He & Lu, 2022) also uses *weak* second-order conditions.

### A.3. On the Optimality of Our Bounds

Revisiting (Carmon et al., 2019a) we can propose a reduction of lower bounds in our setting to lower bounds obtained in that paper. The construction is as follows.

**Class of problems:** Minimizing functions satisfying our Assumption 3.1 on a convex set $\bar{\mathsf{X}}$.

**A particular worst-case problem in this class:** Minimizing the same worst-case objective as in (Carmon et al., 2019a) with additional constraint by a very large ball so that unconstrained stationary point lies in the interior of this ball. No linear constraints.

By our Remark 3.2 this is a consistent situation since their function has Lipschitz gradient and, hence, belongs to our class since the feasible set is compact.

**Class of algorithms:** The same as in (Carmon et al., 2019a).

Then, by the result of (Carmon et al., 2019a) we have that for any algorithm from their rather general class it holds that $\|\nabla f(x_k)\| > \varepsilon$ if $k \leq T = \Theta(\varepsilon^{-2})$. Since our feasible set $\bar{\mathsf{X}}$ is a sufficiently large ball, we have that $x = x_k - \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|} \in \bar{\mathsf{X}}$. Hence $\langle \nabla f(x_k), x - x_k \rangle = -\|\nabla f(x_k)\| < -\varepsilon$. Thus, (11) in our paper does not hold and we have that $x_k$ is not an $\varepsilon$-KKT point at least while $k \leq T = \Theta(\varepsilon^{-2})$. The argument for second-order methods follows the same lines.

Based on this reduction argument, we can conjecture that our bounds are optimal. A formal detailed proof is left for future works.

| | Approach | Constraint | "Optimal" complexity | No proj. | Feasible iterates | Objective | Non-diff. |
|---|---|---|---|---|---|---|---|
| (Ghadimi et al., 2016)<br>(Bogolubsky et al., 2016)<br>(Cartis et al., 2012b)<br>(Birgin & Martínez, 2018)<br>(Cartis et al., 2018) | Proximal | Simple | √ | × | √ | General | × |
| (Andreani et al., 2019)<br>(Andreani et al., 2021) | Augm. Lagrang. | Only cones | × | √ | × | General | × |
| (Lu & Yuan, 2007) | IP | Only cones | × | √ | √ | Quadratic | × |
| (Bian et al., 2015) | IP | Box | √ | √ | √ | Structured | √ |
| (Tseng et al., 2011)<br>(Bomze et al., 2019) | IP | Only $\mathbb{R}^n_+$ | ? | √ | √ | Quadratic | × |
| (Haeser et al., 2019)<br>(O'Neill & Wright, 2020) | IP | Only $\mathbb{R}^n_+$ | √ | √ | √ | General | √ |
| (Dvurechensky & Staudigl, 2024)<br>(He & Lu, 2022) | IP | Only cones | √ | √ | √ | General | √ |
| This paper | IP | **General!** | √ | √ | √ | General | √ |

*Table 1.* Summary of literature. IP stands for Interior-point. "No proj." means that the algorithm does not need to project onto $\bar{\mathsf{X}}$. "Non-diff." means that the objective may be non-differentiable at the relative boundary of the feasible set.

## B. Auxiliary Facts on SCBs

The following properties are taken from (Nesterov, 2018), Lemma 5.4.3, Theorem 5.3.7.

**Proposition B.1.** *Let $h \in \mathcal{H}_\nu(\mathsf{K})$, $x \in \mathsf{K}$, $t > 0$ and $H(x) = \nabla^2 h(x)$. Then,*

$$\langle \nabla h(x), [H(x)]^{-1} \nabla h(x) \rangle \leq \nu. \tag{43}$$

$$\langle \nabla h(x), y - x \rangle < \nu \qquad \forall y \in \mathsf{K}. \tag{44}$$

Note that (44) means that $\nabla h(x) \in \mathsf{NC}_{\mathsf{K}}^{\nu}(x)$. The following fact may be derived from (44) and is its appriximate counterpart.

**Proposition B.2** (Proposition 2.7 (Monteiro et al., 2015))**.** *Let $h \in \mathcal{H}_\nu(\mathsf{K})$, $x \in \mathsf{K}$, and $s \in \mathbb{E}$ satisfy $\|s - \nabla h(x)\|_x^* \leq \xi < 1$. Then, for all $y \in \bar{\mathsf{K}}$,*

$$\langle s, y - x \rangle \leq \nu + \frac{\sqrt{\nu} + \xi}{1 - \xi} \xi. \tag{45}$$

## C. Proof of Theorem 2.1

Let $x^*$ be a local solution for problem (Opt). We consider the following perturbed version of problem (Opt), for which $x^*$ is the unique global solution when $\delta > 0$ sufficiently small,

$$\min_x f(x) + \frac{1}{4}\|x - x^*\|^4 \quad \text{s.t.: } \mathbf{A}x = b,\ x \in \bar{\mathsf{K}},\ \|x - x^*\|^2 \leq \delta. \tag{46}$$

Next, using the barrier $h$ for $\mathsf{K}$, we change the constraint $x \in \mathsf{K}$ to the penalty $\mu_k h(x)$, where $\mu_k > 0$, $\mu_k \to 0$ is a given sequence. This leads us to the following parametric sequence of problems for $k \geq 0$

$$\begin{aligned} &\min\nolimits_x f_k(x) \triangleq f(x) + \tfrac{1}{4}\|x - x^*\|^4 + \mu_k h(x) \\ \text{s.t.:}\quad &\mathbf{A}x = b,\ x \in \operatorname{int}(\mathsf{K}),\ \|x - x^*\|^2 \leq \delta. \end{aligned} \tag{47}$$

From the classical theory of interior penalty methods (Fiacco & McCormick, 1968), it is known that a global solution $x^k$ exists for this problem for all $k$ and that cluster points of $x^k$ are global solutions of (46). Clearly, $\mathbf{A}x^k = b$ and

$x^k \in \mathsf{K} = \operatorname{int}(\mathsf{K}) = \operatorname{int}(\bar{\mathsf{K}})$. Since $\|x^k - x^*\|^2 \le \delta$, the sequence $x^k$ is bounded and, thus, $x^k \to x^*$. This finishes the proof of item 1 of Theorem 2.1.

Since $x^k \to x^*$, for large enough $k$, $x^k$ is a local solution to the problem

$$\min_x f(x) + \frac{1}{4}\|x - x^*\|^4 + \mu_k h(x) \quad \text{s.t.: } \mathbf{A}x = b \tag{48}$$

and $x^k \in \mathsf{K}$. By Assumption 1.1, the system of constraints $\mathbf{A}x = b$ has full rank. Hence, we can write necessary optimality conditions for problem (48) which say that there exists a Lagrange multiplier $y^k \in \mathbb{R}^m$ such that

$$0 = \nabla f(x^k) + \|x^k - x^*\|^2(x^k - x^*) - \mathbf{A}y^k + \mu_k \nabla h(x^k). \tag{49}$$

Let us choose the vector $s^k = -\mu_k \nabla h(x^k)$. Then, item 2 of Theorem 2.1 follows from (49) since $\|x^k - x^*\|^2(x^k - x^*) \to 0$. Further, by (44) with $x = x^k$, we have for any $y \in \bar{\mathsf{K}}$

$$\langle \nabla h(x^k), y - x^k \rangle \le \nu \Leftrightarrow \langle -s^k/\mu_k, y - x^k \rangle \le \nu \Leftrightarrow \langle -s^k, y - x^k \rangle \le \mu_k \nu.$$

Taking $\sigma_k = \mu_k \nu \to 0$, we see that $-s^k \in \mathsf{NC}_{\mathsf{K}}^{\sigma_k}(x^k)$. This proves item 3 of Theorem 2.1.

The second-order differentiability assumption and the full rank condition give the following second-order necessary optimality condition for (48). For all $d \in \mathbb{E}$ such that $\mathbf{A}d = 0$, it holds that

$$\langle (\nabla^2 f(x^k) + \mu_k H(x^k) + \Sigma_k)d, d \rangle \ge 0, \tag{50}$$

where $\Sigma_k = 2(x^k - x^*)(x^k - x^*)^\top + \|x^k - x^*\|^2 \mathbf{I}$, $\mathbf{I}$ being the identity operator. Setting $\theta_k = \mu_k$ and $\delta_k$ as the largest eigenvalue of the positive semi-definite matrix $\Sigma_k$, we conclude that $\theta_k \to 0$ and $\delta_k \to 0$ as $k \to \infty$. This finishes the proof of eq. (6) and Theorem 2.1. □

## D. Missing Proofs from Section 3

### D.1. Proof of Theorem 3.3

Our proof proceeds in four steps. First, we show that $\mathbf{FOABM}(\mu, \varepsilon, L_0, x^0)$ produces points in $\mathsf{X}$, and, thus, is indeed an interior-point method. Then, we proceed to show that the line-search process of finding appropriate $L_k$'s in each iteration is finite, and estimate the total number of attempts in this process. After that, we prove that if the stopping criterion does not hold at iteration $k$, i.e., $\|v^k\|_{x^k} \ge \frac{\varepsilon}{3\nu}$, then the objective $f$ is decreased by a quantity $O(\varepsilon^2)$. From the global lower boundedness of the objective, we derive that the method stops in at most $O(\varepsilon^{-2})$ iterations. Finally, we show the opposite, i.e., that when the stopping criterion holds, the method has generated an $\varepsilon$-KKT point according to Definition 2.2.

#### D.1.1. Interior-Point Property of the Iterates

We start the induction argument by observing that $x^0 \in \mathsf{X}$ by construction. Further, let $x^k \in \mathsf{X} = \mathsf{K} \cap \mathsf{L}$ be the $k$-th iterate of the algorithm with the corresponding step direction $v^k \triangleq v_\mu(x^k)$. By eq. (24), the step-size $\alpha_k$ satisfies $\alpha_k \le \frac{1}{2\|v^k\|_{x^k}}$, and, hence, $\alpha_k \|v^k\|_{x^k} \le 1/2$ for all $k \ge 0$. Thus, by Lemma 1.3, we have $x^{k+1} = x^k + \alpha_k v^k \in \mathsf{K}$. Since, by (23), $\mathbf{A}v^k = 0$, we have that $x^{k+1} \in \mathsf{L}$. Thus, $x^{k+1} \in \mathsf{K} \cap \mathsf{L} = \mathsf{X}$. By induction, we conclude that $(x^k)_{k \ge 0} \subset \mathsf{X}$.

#### D.1.2. Bounding the Number of Backtracking Steps

Consider iteration $k$. The sequence $2^{i_k} L_k$ is increasing as $i_k$ is increasing. Hence, by Assumption 3.1, we know that when $2^{i_k} L_k \ge \max\{M, L_k\}$, the line-search process for sure stops since inequality (25) holds. Thus, $2^{i_k} L_k \le 2\max\{M, L_k\}$ must be the case, and, consequently, $L_{k+1} = 2^{i_k - 1} L_k \le \max\{M, L_k\}$, which, by induction, gives $L_{k+1} \le \bar{M} \triangleq \max\{M, L_0\}$. At the same time, $\log_2\left(\frac{L_{k+1}}{L_k}\right) = i_k - 1, \forall k \ge 0$. Let $N(k)$ denote the number of inner line-search iterations up to the $k-$th iteration of $\mathbf{FOABM}(\mu, \varepsilon, L_0, x^0)$. Then, using that $L_{k+1} \le \bar{M} = \max\{M, L_0\}$, we obtain

$$N(k) = \sum_{j=0}^{k}(i_j + 1) = \sum_{j=0}^{k}(\log_2(L_{j+1}/L_j) + 2) \le 2(k+1) + \max\{\log_2(M/L_0), 0\}.$$

As we see, on average the inner loop ends after two trials.

### D.1.3. Bound for the Number of Outer Iterations

Our goal now is to establish per-iteration decrease of the potential $F_\mu$ if the stopping condition does not yet hold, i.e., $\|v^k\|_{x^k} \geq \frac{\varepsilon}{3\nu}$, and derive from that a global iteration complexity bound for **FOABM**. Let us fix iteration counter $k$. Since $L_{k+1} = 2^{i_k-1}L_k$, the step-size (24) is equivalent to $\alpha_k = \min\left\{\frac{1}{2L_{k+1}+2\mu}, \frac{1}{2\|v^k\|_{x^k}}\right\}$. Hence, $\alpha_k\|v^k\|_{x^k} \leq 1/2$, and (22) with the substitution $t = \alpha_k = \mathtt{t}_{\mu,2L_{k+1}}(x^k)$, $M = 2L_{k+1}$, $x = x^k$, $v_\mu(x^k) \triangleq v^k$ gives:

$$F_\mu(x^{k+1}) - F_\mu(x^k) \leq -\alpha_k\|v^k\|^2_{x^k}\left(1-(L_{k+1}+\mu)\alpha_k\right) \leq -\frac{\alpha_k\|v^k\|^2_{x^k}}{2}, \tag{51}$$

where the last inequality is since $\alpha_k \leq \frac{1}{2(L_{k+1}+\mu)}$. Substituting into (51) the two possible values of the step-size $\alpha_k$ in (24) gives

$$F_\mu(x^{k+1}) - F_\mu(x^k) \leq \begin{cases} -\frac{\|v^k\|^2_{x^k}}{4(L_{k+1}+\mu)} & \text{if}\alpha_k = \frac{1}{2(L_{k+1}+\mu)}, \\ -\frac{\|v^k\|_{x^k}}{4} & \text{if } \alpha_k = \frac{1}{2\|v^k\|_{x^k}}. \end{cases} \tag{52}$$

As we proved in Section D.1.2, $L_{k+1} \leq \bar{M}$. Thus, we obtain that

$$F_\mu(x^{k+1}) - F_\mu(x^k) \leq -\frac{\|v^k\|_{x^k}}{4}\min\left\{1, \frac{\|v^k\|_{x^k}}{\bar{M}+\mu}\right\} \triangleq -\delta_k. \tag{53}$$

Rearranging and summing these inequalities for $k$ from 0 to $K-1$ gives

$$\begin{aligned} K \min_{k=0\dots,K-1}\delta_k &\leq \sum_{k=0}^{K-1}\delta_k \leq F_\mu(x^0) - F_\mu(x^K) \\ &\overset{(1)}{=} f(x^0) - f(x^K) + \mu(h(x^0) - h(x^K)) \leq f(x^0) - f_{\min}(\mathsf{X}) + \varepsilon, \end{aligned} \tag{54}$$

where in the last inequality we used that, by the assumptions of Theorem 3.3, $x^0$ is a $\nu$-analytic center defined in (18) and $\mu = \varepsilon/\nu$, implying that $h(x^0) - h(x^K) \leq \nu = \varepsilon/\mu$. Thus, up to passing to a subsequence, $\delta_k \to 0$, and consequently $\|v^k\|_{x^k} \to 0$ as $k \to \infty$. Hence, the stopping criterion in Algorithm 1 is achievable and the algorithm is correctly defined in this respect.

Let us now assume that the stopping criterion $\|v^k\|_{x^k} < \frac{\varepsilon}{3\nu}$ does not hold for $K$ iterations of **FOABM**. Then, for all $k = 0,\dots,K-1$, we have $\delta_k \geq \min\left\{\frac{\varepsilon}{12\nu}, \frac{\varepsilon^2}{36\nu^2(\bar{M}+\mu)}\right\}$. Using that we set $\mu = \frac{\varepsilon}{\nu}$, it follows from (54) that

$$K\frac{\varepsilon^2}{36\nu^2(\bar{M}+\varepsilon/\nu)} = K\min\left\{\frac{\varepsilon}{12\nu}, \frac{\varepsilon^2}{36\nu^2(\bar{M}+\varepsilon/\nu)}\right\} \leq f(x^0) - f_{\min}(\mathsf{X}) + \varepsilon.$$

Hence, recalling that $\bar{M} = \max\{M, L_0\}$, we obtain

$$K \leq 36(f(x^0) - f_{\min}(\mathsf{X}) + \varepsilon)\cdot\frac{\nu^2(\max\{M, L_0\} + \varepsilon/\nu)}{\varepsilon^2}.$$

Thus, we obtain the bound on the number of iterations on which the stopping criterion is not satisfied. This, combined with the bound for the number of inner steps in Section D.1.2, proves the complexity bound statement of Theorem 3.3.

### D.1.4. Generating $\varepsilon$-KKT Point

To finish the proof of Theorem 3.3, we now show that when Algorithm 1 stops for the first time, it returns a $2\varepsilon$-KKT point of (Opt) according to Definition 2.2.

Clearly, (10) in Definition 2.2 holds by the construction of the algorithm. Thus, we focus on showing (11). Let the stopping criterion hold at iteration $k$, i.e., $\|v^k\|_{x^k} < \frac{\varepsilon}{3\nu}$. Using the optimality condition (20) at iteration $k$ and the definition of the potential (1), we get

$$\nabla f(x^k) - \mathbf{A}^*y^k + \mu\nabla h(x^k) = -H(x^k)v^k \Leftrightarrow [H(x^k)]^{-1}\left(\nabla f(x^k) - \mathbf{A}^*y^k + \mu\nabla h(x^k)\right) = -v^k. \tag{55}$$

Multiplying both equations, taking the square root, and using the stopping criterion $\|v^k\|_{x^k} < \frac{\varepsilon}{3\nu}$ we obtain

$$\|\nabla f(x^k) - \mathbf{A}^* y^k + \mu \nabla h(x^k)\|^*_{x^k} = \|v^k\|_{x^k} < \frac{\varepsilon}{3\nu}, \tag{56}$$

whence, dividing by $\mu$

$$\|-\frac{1}{\mu}(\nabla f(x^k) - \mathbf{A}^* y^k) - \nabla h(x^k)\|^*_{x^k} < \frac{\varepsilon}{3\mu\nu} = \frac{1}{3}. \tag{57}$$

Applying (45) with $s = -\frac{1}{\mu}(\nabla f(x^k) - \mathbf{A}^* y^k)$ and $\xi = \frac{\varepsilon}{3\mu\nu} = \frac{1}{3}$, we obtain

$$\langle -\frac{1}{\mu}(\nabla f(x^k) - \mathbf{A}^* y^k), x - x_k\rangle < \nu + \frac{\sqrt{\nu}+\xi}{1-\xi}\xi = \nu + \frac{\sqrt{\nu}+1/3}{2} \qquad \forall x \in \bar{\mathsf{K}}, \tag{58}$$

whence,

$$\langle \nabla f(x^k) - \mathbf{A}^* y^k, x - x_k\rangle > -\mu\nu - \mu\frac{\sqrt{\nu}+1/3}{2} > -2\varepsilon \qquad \forall x \in \bar{\mathsf{K}}, \tag{59}$$

where we used that $\mu = \frac{\varepsilon}{\nu}$ and that $\nu \geq 1$. Thus, we obtain that (11) holds, which finishes the proof of Theorem 3.3.

# E. Missing Proofs from Section 4

## E.1. Proofs of Preliminary Results

**Proof of the implication (28) ⇒ (27).** We have

$$\|\nabla f(x+v) - \nabla f(x) - \nabla^2 f(x)v\|^*_x = \|\int_0^1 (\nabla^2 f(x+tv) - \nabla^2 f(x))v\, dt\|^*_x$$
$$\leq \int_0^1 \|\nabla^2 f(x+tv) - \nabla^2 f(x)\|_{\mathrm{op},x} \cdot \|v\|_x \, dt \leq \frac{M}{2}\|v\|_x^2.$$

**Proof of (29).** To obtain (29), observe that for all $x \in \mathsf{X}$ and $v \in \mathcal{T}_x$, we have

$$|f(x+v) - f(x) - \langle \nabla f(x), v\rangle - \frac{1}{2}\langle \nabla^2 f(x)v, v\rangle| = |\int_0^1 \langle \nabla f(x+tv) - \nabla f(x) - \frac{1}{2}\nabla^2 f(x)v, v\rangle\, dt|$$
$$\leq \int_0^1 \|\nabla f(x+tv) - \nabla f(x) - \frac{1}{2}\nabla^2 f(x)v\|^*_x \, dt \cdot \|v\|_x \leq \frac{M}{6}\|v\|_x^3.$$

**Proof of Proposition 4.3.** The high-level idea is that after a transition to the basis induced by the affine subspace $\mathsf{L}_0$, the subproblem (32) becomes an unconstrained minimization problem similar to the Cubic Newton step in (Nesterov & Polyak, 2006). To that end, let $\{z_1, \ldots, z_p\}$ be an orthonormal basis of $\mathsf{L}_0$ and the linear operator $\mathbf{Z} : \mathbb{R}^p \to \mathsf{L}_0$ be defined by $\mathbf{Z}w = \sum_{i=1}^p z_i w^i$ for all $w = [w^1; \ldots; w^p]^\top \in \mathbb{R}^p$. Based on this linear map, we define the projected data

$$\mathbf{g} \triangleq \mathbf{Z}^* \nabla F_\mu(x),\ \mathbf{J} \triangleq \mathbf{Z}^* \nabla^2 f(x)\mathbf{Z},\ \mathbf{H} \triangleq \mathbf{Z}^* H(x)\mathbf{Z} \succ 0 \tag{60}$$

and apply it, together with the change of variables $v = \mathbf{Z}u$, to reformulate the search-direction finding problem (32) as an unconstrained cubic-regularized subproblem of finding $u_L \in \mathbb{R}^p$ s.t.

$$u_L \in \operatorname*{Argmin}_{u \in \mathbb{R}^p}\{\langle \mathbf{g}, u\rangle + \frac{1}{2}\langle \mathbf{J}u, u\rangle + \frac{L}{6}\|u\|^3_{\mathbf{H}}\}, \tag{61}$$

where $\|\cdot\|_{\mathbf{H}}$ is the norm induced by the operator $\mathbf{H}$. From (Nesterov & Polyak, 2006), Thm. 10 we deduce

$$\mathbf{J} + \frac{L\|u_L\|_{\mathbf{H}}}{2}\mathbf{H} \succeq 0.$$

Denoting $v_{\mu,L}(x) = \mathbf{Z}u_L$, we see

$$\|u_L\|_{\mathbf{H}} = \langle \mathbf{Z}^* H(x)\mathbf{Z}u_L, u_L\rangle^{1/2} = \langle H(x)(\mathbf{Z}u_L), \mathbf{Z}u_L\rangle^{1/2} = \|v_{\mu,L}(x)\|_x, \text{ and}$$
$$\mathbf{Z}^*\left(\nabla^2 f(x) + \frac{L}{2}\|v_{\mu,L}(x)\|_x H(x)\right)\mathbf{Z} \succeq 0,$$

which implies $\nabla^2 f(x) + \frac{L}{2}\|v_{\mu,L}(x)\|_x H(x) \succeq 0$ over the nullspace $\mathsf{L}_0 = \{v \in \mathbb{E} : \mathbf{A}v = 0\}$. □

The above derivations give us a hint on possible approaches to numerically solve problem (32) in practice. Before the start of the algorithm, as a preprocessing step, we once calculate matrix $\mathbf{Z}$ and use it during the whole algorithm execution. At each iteration, we calculate the new data using (60) and get a standard *unconstrained* cubic subproblem (61). (Nesterov & Polyak, 2006) show how such problems can be transformed to a *convex* problem to which fast convex programming methods could in principle be applied. However, we can also solve it via recent efficient methods based on Lanczos' method (Cartis et al., 2011; Jia et al., 2022). In any case, we can recover our step direction $v_{\mu,L}(x)$ by the matrix vector product $\mathbf{Z}u_L$, where $u_L$ is the solution obtained from this subroutine.

**Derivation of the step-size of Algorithm 2 and derivation of (36).** Our goal now is to construct an admissible step-size policy, given the step direction $v_{\mu,L}(x)$. We act in a similar fashion as in the analysis of the first-order algorithm by applying optimality conditions and estimating the per-iteration decrease of the potential depending on the step-size. Let $x \in \mathsf{X}$ be the current position of the algorithm. Define $x^+(t) \triangleq x + tv_{\mu,L}(x)$, where $t \geq 0$ is a step-size. By Lemma 1.3 and since $v_{\mu,L}(x) \in \mathsf{L}_0$ by (34), we know that $x^+(t)$ is in $\mathsf{X}$ provided that $t \in I_{x,\mu,L} \triangleq [0, \frac{1}{\|v_{\mu,L}(x)\|_x})$. For all such $t$, by (30), we get

$$\begin{aligned} F_\mu(x^+(t)) \leq F_\mu(x) &+ t\langle \nabla F_\mu(x), v_{\mu,L}(x)\rangle + \frac{t^2}{2}\langle \nabla^2 f(x) v_{\mu,L}(x), v_{\mu,L}(x)\rangle \\ &+ \frac{Mt^3}{6}\|v_{\mu,L}(x)\|_x^3 + \mu t^2 \|v_{\mu,L}(x)\|_x^2 \omega(t\|v_{\mu,L}(x)\|_x). \end{aligned} \tag{62}$$

Since $v_{\mu,L}(x) \in \mathsf{L}_0 = \{v \in \mathbb{E} | \mathbf{A}v = 0\}$, multiplying (35) with $v_{\mu,L}(x)$ from the left and the right, and multiplying (33) by $v_{\mu,L}(x)$ and combining with (34), we obtain

$$\langle \nabla^2 f(x) v_{\mu,L}(x), v_{\mu,L}(x)\rangle \geq -\frac{L}{2}\|v_{\mu,L}(x)\|_x^3, \tag{63}$$

$$\langle \nabla F_\mu(x), v_{\mu,L}(x)\rangle + \langle \nabla^2 f(x) v_{\mu,L}(x), v_{\mu,L}(x)\rangle + \frac{L}{2}\|v_{\mu,L}(x)\|_x^3 = 0. \tag{64}$$

Under the additional assumption that $t \leq 2$ and $L \geq M$, we obtain

$$\begin{aligned} & t\langle \nabla F_\mu(x), v_{\mu,L}(x)\rangle + \frac{t^2}{2}\langle \nabla^2 f(x) v_{\mu,L}(x), v_{\mu,L}(x)\rangle + \frac{Mt^3}{6}\|v_{\mu,L}(x)\|_x^3 \\ & \overset{(64)}{=} -t\left(\langle \nabla^2 f(x) v_{\mu,L}(x), v_{\mu,L}(x)\rangle + \frac{L}{2}\|v_{\mu,L}(x)\|_x^3\right) \\ & + \frac{t^2}{2}\langle \nabla^2 f(x) v_{\mu,L}(x), v_{\mu,L}(x)\rangle + \frac{Mt^3}{6}\|v_{\mu,L}(x)\|_x^3 \\ & = \left(\frac{t^2}{2} - t\right)\langle \nabla^2 f(x) v_{\mu,L}(x), v_{\mu,L}(x)\rangle - \frac{Lt}{2}\|v_{\mu,L}(x)\|_x^3 + \frac{Mt^3}{6}\|v_{\mu,L}(x)\|_x^3 \\ & \overset{(63),t\leq 2}{\leq} \left(\frac{t^2}{2} - t\right)\left(-\frac{L}{2}\|v_{\mu,L}(x)\|_x^3\right) - \frac{Lt}{2}\|v_{\mu,L}(x)\|_x^3 + \frac{Mt^3}{6}\|v_{\mu,L}(x)\|_x^3 \\ & = -\|v_{\mu,L}(x)\|_x^3\left(\frac{Lt^2}{4} - \frac{Mt^3}{6}\right) \overset{L\geq M}{\leq} -\|v_{\mu,L}(x)\|_x^3 \frac{Lt^2}{12}(3-2t). \end{aligned}$$

Substituting this into (62), we arrive at

$$\begin{aligned} F_\mu(x^+(t)) &\leq F_\mu(x) - \|v_{\mu,L}(x)\|_x^3 \frac{Lt^2}{12}(3-2t) + \mu t^2\|v_{\mu,L}(x)\|_x^2 \omega(t\|v_{\mu,L}(x)\|_x) \\ &\overset{(4)}{\leq} F_\mu(x) - \|v_{\mu,L}(x)\|_x^3 \frac{Lt^2}{12}(3-2t) + \mu\frac{t^2\|v_{\mu,L}(x)\|_x^2}{2(1-t\|v_{\mu,L}(x)\|_x)}. \end{aligned}$$

for all $t \in I_{x,\mu,L}$. Therefore, if $t\|v_{\mu,L}(x)\|_x \leq 1/2$, we finally obtain

$$\begin{aligned} F_\mu(x^+(t)) - F_\mu(x) &\leq -\frac{Lt^2\|v_{\mu,L}(x)\|_x^3}{12}(3-2t) + \mu t^2\|v_{\mu,L}(x)\|_x^2 \\ &= -\|v_{\mu,L}(x)\|_x^3 \frac{Lt^2}{12}\left(3 - 2t - \frac{12\mu}{L\|v_{\mu,L}(x)\|_x}\right) \triangleq -\eta_x(t). \end{aligned} \tag{65}$$

This is exactly the bound (36). Unfortunately, finding and using the explicit maximizer of $\eta_x(t)$ is quite challenging. But, as we will see, the following step-size is a good and simple alternative:

$$\mathtt{t}_{\mu,L}(x) \triangleq \frac{1}{\max\{1, 2\|v_{\mu,L}(x)\|_x\}} = \min\left\{1, \frac{1}{2\|v_{\mu,L}(x)\|_x}\right\}. \tag{66}$$

Note that $\mathtt{t}_{\mu,L}(x) \leq 1$ and $\mathtt{t}_{\mu,L}(x)\|v_{\mu,L}(x)\|_x \leq 1/2$. Thus, this choice of the step-size is feasible to derive (65).

## E.2. Proof of Theorem 4.4

The main steps of the proof are similar to the analysis of Algorithm 1. We start by showing the feasibility of the iterates and the correctness of the backtracking line-search process, i.e., that this process is finite. We also estimate the total number of attempts in this process. After that, we analyze the per-iteration decrease of $F_\mu$ and show that if the stopping criterion does not hold at iteration $k$, then the objective function is decreased by the value $O(\varepsilon^{3/2})$. This, by the global lower boundedness of the objective, allows us to conclude that the algorithm stops in $O(\varepsilon^{-3/2})$ iterations. Finally, we show the opposite, i.e., that when the stopping criterion holds, the method has generated an approximate second-order KKT point in the sense of Definition 2.3.

### E.2.1. Interior-Point Property of the Iterates

We start the induction argument by observing that $x^0 \in \mathsf{X}$ by construction. Further, let $x^k \in \mathsf{X} = \mathsf{K} \cap \mathsf{L}$ be the $k$-th iterate of the algorithm with the corresponding step direction $v^k \triangleq v_{\mu,L}(x^k)$. By (39), the step-size $\alpha_k$ satisfies $\alpha_k \leq \frac{1}{2\|v^k\|_{x^k}}$. Consequently, $\alpha_k\|v^k\|_{x^k} \leq 1/2$ for all $k \geq 0$, and using Lemma 1.3 as well as equality $\mathbf{A}v^k = 0$ by (38), we have that $x^{k+1} = x^k + \alpha_k v^k \in \mathsf{K} \cap \mathsf{L} = \mathsf{X}$. By induction, it follows that $x^k \in \mathsf{X}$ for all $k \geq 0$.

### E.2.2. Bounding the Number of Backtracking Steps

To bound the number of cycles involved in the line-search process for finding appropriate constants $L_k$, we proceed as in Section D.1.2. Let us fix an iteration $k$. The sequence $L_k = 2^{i_k}M_k$ is increasing as $i_k$ is increasing, and Assumption 4.1 holds. This implies (29), and thus when $L_k = 2^{i_k}M_k \geq \max\{M, M_k\}$, the line-search process for sure stops since inequalities (40) and (41) hold. Hence, $L_k = 2^{i_k}M_k \leq 2\max\{M, M_k\}$ must be the case, and, consequently, $M_{k+1} = \max\{L_k/2, \underline{L}\} \leq \max\{\max\{M, M_k\}, \underline{L}\} = \max\{M, M_k\}$, which, by induction, gives $M_k \leq \bar{M} \triangleq \max\{M, M_0\}$ and $L_k \leq 2\bar{M}$. At the same time, by construction, $M_{k+1} = \max\{2^{i_k - 1}M_k, \underline{L}\} = \max\{L_k/2, \underline{L}\} \geq L_k/2$. Hence, $L_{k+1} = 2^{i_{k+1}}M_{k+1} \geq 2^{i_{k+1}-1}L_k$ and therefore $\log_2\left(\frac{L_{k+1}}{L_k}\right) \geq i_{k+1} - 1, \forall k \geq 0$. At the same time, at iteration 0 we have $L_0 = 2^{i_0}M_0 \leq 2\bar{M}$, whence, $i_0 \leq \log_2\left(\frac{2\bar{M}}{M_0}\right)$. Let $N(k)$ denote the number of inner line-search iterations up to iteration $k$ of **SOABM**. Then,

$$N(k) = \sum_{j=0}^{k}(i_j + 1) \leq i_0 + 1 + \sum_{j=1}^{k}\left(\log_2\left(\frac{L_j}{L_{j-1}}\right) + 2\right) \leq 2(k+1) + 2\log_2\left(\frac{2\bar{M}}{M_0}\right),$$

since $L_k \leq 2\bar{M} = 2\max\{M, M_0\}$ in the last step. Thus, on average, the inner loop ends after two trials.

### E.2.3. Bound for the Number of Outer Iterations

Our goal now is to establish per-iteration decrease of the potential $F_\mu$ if the stopping condition does not yet hold, and derive from that a global iteration complexity bound for **SOABM**. Let us fix iteration counter $k$. As said, the main assumption of this subsection is that the stopping criterion is not satisfied, i.e., either $\|v^k\|_{x^k} \geq \Delta_k$ or $\|v^{k-1}\|_{x^{k-1}} \geq \Delta_{k-1}$. Without loss of generality, we assume that the first inequality holds, i.e., $\|v^k\|_{x^k} \geq \Delta_k$, and consider iteration $k$. Otherwise, if the second inequality holds, the same derivations can be made considering the iteration $k-1$ and using the second inequality $\|v^{k-1}\|_{x^{k-1}} \geq \Delta_{k-1}$. Thus, at the end of the $k$-th iteration

$$\|v^k\|_{x^k} \geq \Delta_k = \sqrt{\frac{\varepsilon}{12L_k\nu}}. \tag{67}$$

Since the step-size $\alpha_k = \min\{1, \frac{1}{2\|v^k\|_{x^k}}\} = \mathtt{t}_{\mu,L_k}(x^k)$ in (39) satisfies $\alpha_k \leq 1$ and $\alpha_k\|v^k\|_{x^k} \leq 1/2$ (see (66) and a remark after it), we can repeat the derivations of (65), changing (29) to (40). In this way we obtain the following counterpart of (65) with $t = \alpha_k$, $L = L_k$, $x = x^k$, $v_{\mu,L_k}(x^k) \triangleq v^k$:

$$F_\mu(x^{k+1}) - F_\mu(x^k) \leq -\|v^k\|_{x^k}^3 \frac{L_k\alpha_k^2}{12}\left(3 - 2\alpha_k - \frac{12\mu}{L_k\|v^k\|_{x^k}}\right) \leq -\|v^k\|_{x^k}^3 \frac{L_k\alpha_k^2}{12}\left(1 - \frac{12\mu}{L_k\|v^k\|_{x^k}}\right), \tag{68}$$

where in the last inequality we used that $\alpha_k \leq 1$ by construction. Substituting $\mu = \frac{\varepsilon}{4\nu}$, and using (67), we obtain

$$1 - \frac{12\mu}{L_k\|v^k\|_{x^k}} = 1 - \frac{12\varepsilon}{4\nu L_k\|v^k\|_{x^k}} \overset{(67)}{\geq} 1 - \frac{3\varepsilon}{\nu L_k\sqrt{\frac{\varepsilon}{12L_k\nu}}} = 1 - \frac{6\sqrt{\varepsilon}}{\sqrt{3\nu L_k}} \geq 1 - \frac{6\sqrt{\varepsilon}}{\sqrt{3\cdot 144\nu\varepsilon}} \geq \frac{1}{2},$$

using that, by construction, $L_k = 2^{i_k}M_k \geq \underline{L} = 144\varepsilon$ and that $\nu \geq 1$. Hence, from (68),

$$F_\mu(x^{k+1}) - F_\mu(x^k) \leq -\|v^k\|_{x^k}^3 \frac{L_k\alpha_k^2}{24}. \tag{69}$$

Substituting into (69) the two possible values of the step-size $\alpha_k$ in (39) gives

$$F_\mu(x^{k+1}) - F_\mu(x^k) \leq \begin{cases} -\|v^k\|_{x^k}^3 \frac{L_k}{24}, & \text{if } \alpha_k = 1, \\ -\|v^k\|_{x^k} \frac{L_k}{96}, & \text{if } \alpha_k = \frac{1}{2\|v^k\|_{x^k}}. \end{cases} \tag{70}$$

This implies

$$F_\mu(x^{k+1}) - F_\mu(x^k) \leq -\frac{L_k\|v^k\|_{x^k}}{96}\min\left\{1, 4\|v^k\|_{x^k}^2\right\} \triangleq -\delta_k. \tag{71}$$

Rearranging and summing these inequalities for $k$ from 0 to $K-1$, and using that $L_k \geq \underline{L}$, we obtain

$$\begin{aligned} K \min_{k=0,\ldots,K-1} \frac{\underline{L}\|v^k\|_{x^k}}{96}\min\left\{1, 4\|v^k\|_{x^k}^2\right\} &\leq \sum_{k=0}^{K-1}\delta_k \leq F_\mu(x^0) - F_\mu(x^K) \\ &\overset{(1)}{=} f(x^0) - f(x^K) + \mu(h(x^0) - h(x^K)) \leq f(x^0) - f_{\min}(\mathsf{X}) + \varepsilon, \end{aligned} \tag{72}$$

where we used that, by the assumptions of Theorem 4.4, $x^0$ is a $4\nu$-analytic center defined in (18) and $\mu = \frac{\varepsilon}{4\nu}$, implying that $h(x^0) - h(x^K) \leq 4\nu = \varepsilon/\mu$. Thus, up to passing to a subsequence, we have $\|v^k\|_{x^k} \to 0$ as $k \to \infty$, which makes the stopping criterion in Algorithm 2 achievable.

Assume now that the stopping criterion does not hold for $K$ iterations of **SOABM**. Then, for all $k = 0, \ldots, K-1$, it holds that

$$\begin{aligned} \delta_k = \frac{L_k}{96}\min\left\{\|v^k\|_{x^k}, 4\|v^k\|_{x^k}^3\right\} &\overset{(67)}{\geq} \frac{L_k}{96}\min\left\{\sqrt{\frac{\varepsilon}{12L_k\nu}}, \frac{4\varepsilon^{3/2}}{12^{3/2}L_k^{3/2}\nu^{3/2}}\right\} \\ &\overset{L_k\leq 2\bar{M}, \nu\geq 1}{\geq} \frac{1}{96}\min\left\{\frac{L_k\sqrt{\varepsilon}}{\sqrt{24\bar{M}}\nu^{3/2}}, \frac{\varepsilon^{3/2}}{2\cdot 3^{3/2}L_k^{1/2}\nu^{3/2}}\right\} \\ &\overset{L_k\leq 2\bar{M}, L_k\geq 144\varepsilon}{\geq} \frac{1}{96}\min\left\{\frac{(144\varepsilon)\cdot\sqrt{\varepsilon}}{\sqrt{24\bar{M}}\nu^{3/2}}, \frac{\varepsilon^{3/2}}{6\sqrt{6\bar{M}}\nu^{3/2}}\right\} = \frac{\varepsilon^{3/2}}{576\nu^{3/2}\sqrt{6\bar{M}}}. \end{aligned} \tag{73}$$

Thus, from (72)

$$K\frac{\varepsilon^{3/2}}{576\nu^{3/2}\sqrt{6\bar{M}}} \leq f(x^0) - f_{\min}(\mathsf{X}) + \varepsilon.$$

Hence, recalling that $\bar{M} = \max\{M_0, M\}$, we obtain $K \leq \frac{576\nu^{3/2}\sqrt{6\max\{M_0,M\}}(f(x^0) - f_{\min}(\mathsf{X}) + \varepsilon)}{\varepsilon^{3/2}}$. Thus, we obtain the bound on the number of iterations on which the stopping criterion is not satisfied. This, combined with the bound for the number of inner steps in Section E.2.2, proves the complexity bound statement of Theorem 4.4.

### E.2.4. Generating $(\varepsilon_1, \varepsilon_2)$-2KKT Point

To finish the proof of Theorem 4.4, we show that if the stopping criterion in Algorithm 2 holds, i.e., $\|v^{k-1}\|_{x^{k-1}} < \Delta_{k-1}$ and $\|v^k\|_{x^k} < \Delta_k$, then the algorithm has generated an $(\varepsilon_1, \varepsilon_2)$-2KKT point of (Opt) according to Definition 2.3, with $\varepsilon_1 = \varepsilon$ and $\varepsilon_2 = \frac{\max\{M_0, M\}\varepsilon}{24\nu}$.

Clearly, (12) in Definition 2.3 holds by the construction of the algorithm. Thus, we focus on showing (13) and (14).

Let the stopping criterion hold at iteration $k$. First, we focus on showing (13). Using the first-order optimality condition (33) for the subproblem (38) solved at iteration $k-1$, there exists a Lagrange multiplier $y^{k-1} \in \mathbb{R}^m$ such that (33) holds. Now, expanding the definition of the potential (1) and adding $\nabla f(x^k)$ to both sides, we obtain from (33)

$$\begin{aligned}\nabla f(x^k) - &\mathbf{A}^* y^{k-1} + \mu \nabla h(x^{k-1}) \\ &= \nabla f(x^k) - \nabla f(x^{k-1}) - \nabla^2 f(x^{k-1}) v^{k-1} - \frac{L_{k-1}}{2}\|v^{k-1}\|_{x^{k-1}} H(x^{k-1}) v^{k-1}.\end{aligned}$$

Setting $s^k \triangleq \nabla f(x^k) - \mathbf{A}^* y^{k-1} \in \mathbb{E}^*$ and $g^{k-1} \triangleq -\mu \nabla h(x^{k-1})$, after multiplication by $[H(x^{k-1})]^{-1}$, this is equivalent to

$$[H(x^{k-1})]^{-1}\left(s^k - g^{k-1}\right) = [H(x^{k-1})]^{-1}\left(\nabla f(x^k) - \nabla f(x^{k-1}) - \nabla^2 f(x^{k-1}) v^{k-1} - \frac{L_{k-1}}{2}\|v^{k-1}\|_{x^{k-1}} H(x^{k-1}) v^{k-1}\right).$$

Multiplying both of the above equalities, we arrive at

$$\left(\|s^k - g^{k-1}\|^*_{x^{k-1}}\right)^2 = \left(\left\|\nabla f(x^k) - \nabla f(x^{k-1}) - \nabla^2 f(x^{k-1}) v^{k-1} - \frac{L_{k-1}}{2}\|v^{k-1}\|_{x^{k-1}} H(x^{k-1}) v^{k-1}\right\|^*_{x^{k-1}}\right)^2.$$

Taking the square root and applying the triangle inequality, we obtain

$$\begin{aligned}\|s^k - g^{k-1}\|^*_{x^{k-1}} &\leq \|\nabla f(x^k) - \nabla f(x^{k-1}) - \nabla^2 f(x^{k-1}) v^{k-1}\|^*_{x^{k-1}} + \frac{L_{k-1}}{2}\|v^{k-1}\|^2_{x^{k-1}} \\ &\overset{(41)}{\leq} \frac{L_{k-1}}{2}\|\alpha_{k-1} v^{k-1}\|^2_{x_{k-1}} + \frac{L_{k-1}}{2}\|v^{k-1}\|^2_{x^{k-1}}.\end{aligned} \tag{74}$$

Since the stopping criterion holds, at iteration $k-1$ we have

$$\|v^{k-1}\|_{x^{k-1}} < \Delta_{k-1} = \sqrt{\frac{\varepsilon}{12 L_{k-1}\nu}} \leq \sqrt{\frac{\varepsilon}{12 \cdot 144\varepsilon\nu}} < \frac{1}{2}, \tag{75}$$

where we used that, by construction, $L_{k-1} \geq \underline{L} = 144\varepsilon$ and that $\nu \geq 1$. Hence, by (39), we have that $\alpha_{k-1} = 1$ and $x^k = x^{k-1} + v^{k-1}$. This, in turn, implies that

$$\|s^k - g^{k-1}\|^*_{x^{k-1}} \overset{(74)}{\leq} L_{k-1}\|v^{k-1}\|^2_{x^{k-1}} \leq L_{k-1}\Delta^2_{k-1} = \frac{\varepsilon}{12\nu}, \tag{76}$$

where we used the stopping criterion

$$\|v^{k-1}\|_{x^{k-1}} < \Delta_{k-1} = \sqrt{\frac{\varepsilon}{12 L_{k-1}\nu}}$$

Recalling that $s^k \triangleq \nabla f(x^k) - \mathbf{A}^* y^{k-1} \in \mathbb{E}^*$ and $g^{k-1} \triangleq -\mu \nabla h(x^{k-1})$, we obtain from (76)

$$\|-\frac{1}{\mu}(\nabla f(x^k) - \mathbf{A}^* y^{k-1}) - \nabla h(x^{k-1})\|^*_{x^{k-1}} \leq \frac{\varepsilon}{12\mu\nu} = \frac{1}{3}, \tag{77}$$

where the last equality uses that $\mu = \frac{\varepsilon}{4\nu}$. Applying (45) with $s = -\frac{1}{\mu}(\nabla f(x^k) - \mathbf{A}^* y^{k-1})$ and $\xi = \frac{\varepsilon}{12\mu\nu} = \frac{1}{3}$, we obtain

$$\langle -\frac{1}{\mu}(\nabla f(x^k) - \mathbf{A}^* y^{k-1}), x - x_k\rangle < \nu + \frac{\sqrt{\nu} + \xi}{1 - \xi}\xi = \nu + \frac{\sqrt{\nu} + 1/3}{2} \qquad \forall x \in \bar{\mathsf{K}}, \tag{78}$$

whence,

$$\langle \nabla f(x^k) - \mathbf{A}^* y^{k-1}, x - x_k \rangle > -\mu\nu - \mu\frac{\sqrt{\nu}+1/3}{2} > -\varepsilon \qquad \forall x \in \bar{\mathsf{K}}, \tag{79}$$

where we used that $\mu = \frac{\varepsilon}{4\nu}$ and that $\nu \geq 1$. Thus, we obtain that (13) holds.
Finally, we show the second-order condition (14). By inequality (35) for subproblem (38) solved at iteration $k$, we obtain on $\mathsf{L}_0$

$$\begin{aligned} \nabla^2 f(x^k) &\succeq -\frac{L_k\|v^k\|_{x^k}}{2}H(x^k) \succeq -\frac{L_k\Delta_k}{2}H(x^k) \\ &= -\frac{L_k}{2}\sqrt{\frac{\varepsilon}{12L_k\nu}}H(x^k) = -\frac{\sqrt{L_k\varepsilon}}{(48\nu)^{1/2}}H(x^k) \succeq -\frac{\sqrt{2\bar{M}\varepsilon}}{(48\nu)^{1/2}}H(x^k) = -\frac{\sqrt{\bar{M}\varepsilon}}{(24\nu)^{1/2}}H(x^k), \end{aligned} \tag{80}$$

where we used the second part of the stopping criterion, i.e., $\|v^k\|_{x^k} < \Delta_k$ and that $L_k \leq 2\bar{M} = 2\max\{M, M_0\}$ (see Section E.2.2). Thus, (14) holds with $\varepsilon_2 = \frac{\max\{M,M_0\}\varepsilon}{24\nu}$, which finishes the proof of Theorem 4.4.