# Improved Analysis of Restarted Accelerated Gradient and Augmented Lagrangian Methods via Inexact Proximal Point Frameworks

Matthew X. Burns* Jiaming Liang †

February 19, 2026

**Abstract**

This paper studies a class of double-loop (inner-outer) algorithms for convex composite optimization. For unconstrained problems, we develop a restarted accelerated composite gradient method that attains the optimal first-order complexity in both the convex and strongly convex settings. For linearly constrained problems, we introduce inexact augmented Lagrangian methods, including a basic method and an outer-accelerated variant, and establish near-optimal first-order complexity for both methods. The established complexity bounds follow from a unified analysis based on new inexact proximal point frameworks that accommodate relative and absolute inexactness, acceleration, and strongly convex objectives. Numerical experiments on LASSO and linearly constrained quadratic programs demonstrate the practical efficiency of the proposed methods.

**Key words.** Convex composite optimization, Accelerated gradient method, Augmented Lagrangian method, Proximal point method, Optimal iteration-complexity

**AMS subject classifications.** 49M37, 65K05, 68Q25, 90C25, 90C30, 90C60

## 1 Introduction

In this paper, we consider two optimization problems: the convex smooth composite optimization (CSCO) problem

$$\phi_* := \min_{x\in\mathbb{R}^n}\{\phi(x) := f(x) + h(x)\}, \tag{1}$$

and the linearly constrained CSCO (LC-CSCO) problem

$$\hat{\phi}_* := \min_{x\in\mathbb{R}^n}\{\phi(x) := f(x) + h(x) : Ax = b\}, \tag{2}$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ define $m \leq n$ linear equality constraints. In both problems, we assume that i) $f, h : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ are closed proper convex functions such that $\operatorname{dom} h \subset \operatorname{dom} f$,

*Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627 (email: mburns13@ur.rochester.edu).

†Goergen Institute for Data Science and Artificial Intelligence (GIDS-AI) and Department of Computer Science, University of Rochester, Rochester, NY 14620 (email: jiaming.liang@rochester.edu). This work was partially supported by AFOSR grant FA9550-25-1-0182.

ii) $f$ is $L_f$-smooth on $\mathbb{R}^n$, and iii) $h$ has a computable proximal mapping. Moreover, we also assume for LC-CSCO that $\operatorname{dom} h$ is bounded with diameter $D$ and Slater's condition is satisfied.

For any $\varepsilon > 0$, we say a point $x \in \mathbb{R}^n$ is an $\varepsilon$-solution to (1) if $\phi(x) - \phi_* \leq \varepsilon$. To define an optimality criterion for (2), we consider the unconstrained primal-dual reformulation,

$$\max_{\lambda \in \mathbb{R}^m} \min_{x \in \mathbb{R}^n} \{\mathcal{L}(x, \lambda) := \phi(x) + \langle \lambda, Ax - b \rangle\},$$

where $\mathcal{L}$ is the Lagrangian function and $\lambda \in \mathbb{R}^m$ is the Lagrange multiplier for the constraint $Ax = b$. Slater's condition implies the strong duality equivalence

$$\hat{\phi}_* = \max_{\lambda \in \mathbb{R}^m} \big\{ d(\lambda) := \min_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda) \big\}, \tag{3}$$

where $d(\lambda)$ is the Lagrangian dual of (2). For fixed $\varepsilon > 0$, we call the pair $(x, \lambda)$ an $\varepsilon$-primal-dual solution to (2) if it is an $\varepsilon$-stationary point of $\mathcal{L}$, that is for some $v \in \mathbb{R}^n$

$$v \in \partial h(x) + \nabla f(x) + A^\top \lambda, \quad \|v\| \leq \varepsilon, \quad \|Ax - b\| \leq \varepsilon. \tag{4}$$

Additionally, we call a point $x \in \operatorname{dom} h$ an $\varepsilon$-primal solution if

$$|\phi(x) - \hat{\phi}_*| \leq \varepsilon, \quad \|Ax - b\| \leq \varepsilon, \tag{5}$$

We can show (see Lemma B.3 in Appendix B) that (4) implies an $\mathcal{O}(\varepsilon)$ primal solution, hence we focus on (4) in this work for generality.

**Literature Review.** Nesterov's accelerated composite gradient (ACG) method is standard for solving CSCO problems. First proposed for purely smooth problems ($h = 0$) [34], accelerated methods have since been extended to the composite setting [3, 5, 30, 35], where they achieve optimal complexity $\mathcal{O}(\varepsilon^{-1/2})$ for obtaining an $\varepsilon$-solution to (1). However, some undesirable phenomena are observed in practice, namely oscillations in the objective value. "Restarted" ACG methods are a widely used strategy to improve ACG performance and suppress oscillations. A restarted ACG method periodically resets the acceleration scheme according to some predefined rule [1, 2, 37, 48]. The seminal work [37] proposed "gradient" and "function value" restart heuristics which restart when the gradient forms an acute angle with the update direction ("gradient") or when the function value increases ("function value"). While these restart criteria are empirically performant, they were initially heuristic strategies without theoretical support. Recent work has shown that gradient restart achieves optimal rates in the strongly convex setting [4], however the authors do not consider the more general class of merely convex objectives. A "speed restart" strategy was further proposed by [48]. Motivated by continuous-time ODE analysis, discrete-time speed restart resets acceleration when $\|x_k - x_{k-1}\| < \|x_{k-1} - x_{k-2}\|$. While a convergence analysis was presented for the continuous-time limit, the final bound contains some constants which are simply shown to exist, lacking exact characterization. Parameter-free restarting schemes for strongly convex optimization were proposed by [49] based on an estimation procedure for the (unknown) strong convexity modulus.

A classical method for solving LC-CSCO problems is the augmented Lagrangian method (ALM), also known as the method of multipliers, which has the iteration

$$x_{k+1} = \operatorname*{argmin}_{u \in \mathbb{R}^n} \mathcal{L}_\rho(u, \lambda_k), \tag{6}$$

$$\lambda_{k+1} = \lambda_k + \rho(Ax_{k+1} - b), \tag{7}$$

where

$$\mathcal{L}_\rho(x, \lambda) = \phi(x) + \langle \lambda, Ax - b \rangle + \frac{\rho}{2} \|Ax - b\|^2 \tag{8}$$

is the augmented Lagrangian with penalty coefficient $\rho > 0$. First analyzed by Hestenes [15] and Powell [41], variants of ALM have become standard methods for linearly-constrained optimization. The classical exact ALM is typically intractable to implement, motivating the development of the inexact ALM (I-ALM),

$$x_{k+1} \approx \operatorname*{argmin}_{u \in \mathbb{R}^n} \mathcal{L}_\rho(u, \lambda_k), \tag{9}$$

$$\lambda_{k+1} = \lambda_k + \rho(Ax_{k+1} - b), \tag{10}$$

which permits some inexactness in the primal minimization step. While analysis of the I-ALM dates back to Rockafellar [42], non-asymptotic guarantees for more general problem classes have emerged only recently. The authors of [18] provided non-ergodic complexity bounds for an I-ALM when $h(x) = \delta_Q(x)$ is the indicator function of a compact convex set $Q$. In a pattern replicated in later works, [18] used ACG as a first-order inner solver for (9) and separated their "inner" and "outer" complexity analyses. The baseline ACG-based I-ALM was shown to have an $\mathcal{O}(\varepsilon^{-7/4})$ complexity. To improve the complexity, [18] added a strongly convex perturbation to (8), $\gamma_p\|x - x_0\|^2/2$ for some $x_0 \in \operatorname{dom} h$ with $\gamma_p \propto \varepsilon/D$ to maintain an $\varepsilon$-primal-dual solution to the original problem. The perturbation seemingly improved I-ALM iteration complexity to $\tilde{\mathcal{O}}(\varepsilon^{-1})$, however, as shown by [26], this perturbation adds a hidden $\varepsilon$ dependence. Accordingly, further studies have rectified and extended non-asymptotic guarantees for the I-ALM. Authors have proven $\tilde{\mathcal{O}}(\varepsilon^{-1})$ complexities by considering ergodic iterates [40, 53] and geometrically increasing penalty terms [26, 53]. Several works have also extended the problem class to include nonlinear inequality constraints [26, 53], general simple nonsmooth $h$ [25, 26, 53], and projection-free inner subroutines [25]. We refer interested readers to the recent survey [9] for a more comprehensive treatment of the ALM in mathematical programming. **Remark.** Numerous works that use inexact first-order subroutines also assume that $\operatorname{dom} h$ is bounded [18, 25, 26, 40, 53]. Interestingly, works which do not assume boundedness also do not rely on inexact first-order subroutines, instead using explicit minimization (either as a theoretical oracle, a linear program, or as a linearized approximation [38, 44, 52]).

While the Restarted ACG and I-ALM algorithms target distinct problem classes, we can view them in a unified perspective by considering inexact proximal point (IPP) methods [42, 43, 45, 47], which solve a generic optimization problem $\Phi_* := \min_{x \in \mathbb{R}^n} \Phi(x)$ by the iteration

$$x_{k+1} \approx \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \Phi(x) + \frac{1}{2\lambda}\|x - x_k\|^2 \right\}.$$

The approximation "$\approx$" can be characterized in a number of ways, either bounded by some absolute tolerance $\delta_k \geq 0$ [43, 45] or by some relative term $\|x_{k+1} - x_k\|$ [31, 32, 45]. IPP frameworks have long been used to analyze algorithms for optimization over problems with convex structures. Rockafellar's absolute error [42] framework has repeatedly been used for I-ALM analysis [25, 53]. Solodov and Svaiter's HPE framework provided the first iteration-complexity bound for ADMM [33] and its accelerated extension [32] has been instrumental in the development of high-order methods [8, 13, 17].

**Contributions.** For CSCO problems, we propose a novel Restarted ACG method (Algorithm 2) that achieves the same optimal iteration-complexity as that of ACG in both convex and strongly convex settings. To our knowledge, this is a novel result in the restarted ACG literature. For LC-CSCO problems, we first prove that a classical I-ALM algorithm (Algorithm 3) achieves near-optimal, non-ergodic $\tilde{\mathcal{O}}(\varepsilon^{-1})$ complexity with a fixed penalty parameter $\rho > 0$. To our knowledge, this is a novel finding in the ALM literature, where previous methods either require geometrically increasing $\rho$ or ergodic convergence to achieve near-optimal complexity. Building on our analysis of

I-ALM, we propose a dual-accelerated inexact "fast" ALM (I-FALM, Algorithm 4), which achieves $\tilde{\mathcal{O}}(\varepsilon^{-1})$ non-ergodic complexity with improved dependence on the domain diameter $D$. Table 1 places the proposed I-ALM and I-FALM methods in the context of the broader I-ALM literature, where the stated complexity is to find an $\varepsilon$-primal solution in the sense of (5). Numerical experiments suggest that Algorithm 2 is competitive with existing restart schemes, as well as showing that Algorithm 4 can significantly outperform existing non-ergodic I-ALM variants.

Furthermore, we introduce two analytical frameworks, namely lower oracle approximation (LOrA) and its accelerated variant, fast LOrA (FLOrA), which provide a unified and principled foundation for analyzing algorithms for solving either CSCO or LC-CSCO. LOrA is an IPP framework built upon "lower estimation" functions that arise naturally in the analysis of convex optimization methods. It encompasses, as special cases, proximal gradient methods, proximal bundle methods, and I-ALM. By incorporating Nesterov's acceleration scheme into LOrA, we develop FLOrA, which further captures ACG, Restarted ACG, and I-FALM as instances.

**Organization.** Section 2 provides an overview of ACG and introduces the Restarted ACG algorithm along with its optimal iteration-complexity. Section 3 provides the setup and complexity bounds for the I-ALM and I-FALM algorithms in Subsections 3.1 and 3.2, respectively. Section 4 introduces the LOrA and FLOrA frameworks along with their theoretical guarantees. Building on the two frameworks, Section 5 proves the main complexity results presented in Sections 2 and 3. Preliminary computational results are reported in Section 6. Section 7 provides concluding remarks and potential future directions. Appendix A provides additional details for numerical experiments. Technical lemmas can be found in Appendix B. LOrA and FLOrA analyses are presented in Appendix C. Appendices D and E contain deferred proofs relevant to Sections 2 and 3, respectively.

| Paper | Alg. | Complexity | $\rho$ | $\phi$ | Constraints | Subroutine | Conv. Pt. |
|---|---|---|---|---|---|---|---|
| [18] | I-ALM | $\mathcal{O}(\varepsilon^{-7/4})$ | Static | $f+\delta_Q$ | Linear | ACG | Non-Erg. |
| [40] | IFAL | $\mathcal{O}(\varepsilon^{-1})$ | Static | $f+\delta_Q$ | Linear | ACG | Erg. |
| [25] | I-ALM | $\mathcal{O}(\varepsilon^{-2})$ | Static | $f+h$ | Linear | ACG/CG | Non-Erg. |
| [53] | I-ALM | $\mathcal{O}(\varepsilon^{-1})$ | Geo. | $f+h$ | Nonlinear | ACG | Non-Erg. |
| | I-ALM | $\mathcal{O}(\varepsilon^{-1})$ | St./Geo. | $f+h$ | Nonlinear | ACG | Erg. |
| [26] | aI-ALM | $\tilde{\mathcal{O}}(\varepsilon^{-1})$ | Geo. | $f+h$ | Nonlinear | ACG | Non-Erg. |
| [21] | LPALM | $\mathcal{O}(\varepsilon^{-1})$ | Static | $f+h$ | Linear | Prox | Non-Erg. |
| **TW** | Alg. 3 | $\tilde{\mathcal{O}}(\varepsilon^{-1})$ | Static | $f+h$ | Linear | ACG | Non-Erg. |
| **TW** | Alg. 4 | $\tilde{\mathcal{O}}(\varepsilon^{-1})$ | Static | $f+h$ | Linear | ACG | Non-Erg. |

Table 1: Non-exhaustive summary of related works on I-ALM. **TW** indicates "This Work". For simplicity, we use the common term "ACG" to refer to either ACG (Algorithm 1) or related variants such as FISTA [5]. "CG" refers to the Conditional Gradient (or Frank-Wolfe) algorithm [7, 11]. "Prox" refers to a single, closed-form proximal mapping for a linearized augmented Lagrangian model [21, 38]. "Static" $\rho$ selection refers to choosing a constant penalty $\rho$ across all iterations, while "Geo(metric)" refers to a geometrically increasing $\rho$, i.e., $\rho_k = \rho_0 \cdot \beta^k$ for some $\beta > 1$. "Conv. Pt." refers to the point of convergence, where "Erg(odic)" refers to convergence in an averaged point (e.g., $\hat{x}_k = k^{-1}\sum_{i=1}^{k} x_i$) while "Non-Erg(odic)" directly shows convergence in some single iterate $x_k$ (e.g., the best or the last). $\delta_Q$ is taken to be the indicator function of some simple, closed convex set $Q$, $h$ is a simple, possibly nonsmooth closed convex function, and $f$ is a smooth closed convex function. Algorithm acronyms are: "IFAL" is "Iterative Fast Augmented Lagrangian", "aI-ALM" is "adaptive I-ALM", and "LPALM" is "Linearized Proximal ALM". All iteration-complexity results are to obtain an $\varepsilon$-primal solution to (2) in the sense of (5).

### 1.1 Basic Definitions and Notation

The set of real numbers is denoted by $\mathbb{R}$, non-negative reals by $\mathbb{R}_+$, and positive reals by $\mathbb{R}_{++}$. Let $\mathbb{R}^n$ be the $n$-dimensional Euclidean space equipped with the standard inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. Let $\mathbb{R}^{n\times n}$ be the space of real-valued $n \times n$ matrices equipped with the spectral norm

$$\|A\| = \sup_{x\in\mathbb{R}^n} \{\|Ax\| : \|x\| \leq 1\}.$$

For a convex set $Q \subseteq \mathbb{R}^n$, we define the *diameter* $D$ as $D = \sup_{x,y\in Q}\{\|x-y\|\}$. If $D < \infty$, then $Q$ is bounded. We define the *relative interior* of $Q$, relint$(Q)$ as

$$\text{relint}(Q) = \{x \in Q : B(x, r) \cap \text{affine}(Q) \subseteq Q \text{ for some } r > 0\},$$

where affine$(Q)$ is the affine hull of $Q$. We say that (2) satisfies *Slater's condition* if there exists a feasible point in relint$(\text{dom}\, h)$, i.e.,

$$\text{relint}(\text{dom}\, h) \cap \{x \in \mathbb{R}^n : Ax = b\} \neq \emptyset. \tag{11}$$

For a proper function $f$, the *subdifferential* of $f$ at $x \in \text{dom}\, f$ is denoted by

$$\partial f(x) := \{s \in \mathbb{R}^n : f(y) \geq f(x) + \langle s, y - x\rangle, \forall y \in \mathbb{R}^n\}.$$

For a given *subgradient* $f'(x) \in \partial f(x)$, we denote the *linearization of $f$ at $x$* by $\ell_f(\cdot; x)$, which is defined as

$$\ell_f(\cdot; x) := f(x) + \langle f'(x), \cdot - x\rangle.$$

For a function $f : \mathbb{R}^n \to (-\infty, +\infty]$, we denote its effective domain by $\text{dom}\, f = \{x : f(x) < +\infty\}$. We say that $f$ is *$\mu$-strongly convex* for some $\mu > 0$ if for every $x, y \in \text{dom}\, f$ and $\lambda \in [0, 1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\lambda(1-\lambda)\mu}{2}\|x - y\|^2.$$

Equivalently, $f$ is $\mu$-strongly convex if for every $x, y \in \text{dom}\, f$ and all $f'(x) \in \partial f(x) \neq \emptyset$,

$$f(y) - f(x) - \langle f'(x), y - x\rangle \geq \frac{\mu}{2}\|x - y\|^2.$$

With $\mu = 0$ we recover the standard definitions of convexity. We denote the set of proper closed $\mu$-strongly convex functions over set $Q$ as $\overline{\underset{\mu}{\text{Conv}}}(Q)$, with $\overline{\text{Conv}}(Q)$ used if $\mu = 0$.

We say that a differentiable function $f$ is *$L_f$-smooth* if $\nabla f$ is $L_f$-Lipschitz continuous on $\mathbb{R}^n$. Equivalently, $f$ is $L_f$-smooth if there exists an $L_f > 0$ such that for every $x, y \in \mathbb{R}^n$

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{L_f}{2}\|x - y\|^2. \tag{12}$$

We define the *proximal mapping* (or "prox mapping") of a closed convex function $h$ as

$$\text{prox}_h(x) = \underset{y\in\mathbb{R}^n}{\text{argmin}} \left\{h(y) + \frac{1}{2}\|x - y\|^2\right\}.$$

We say $h$ is *simple* if it has a computable prox mapping. We define the *iteration-complexity* of an algorithm as the number of prox mappings it needs to solve a problem to a specified tolerance.

Given a positive scalar $\lambda$ and a composite function $\phi(x) = f(x) + h(x)$, where $f$ is smooth and $h$ has an available prox mapping, we define the *gradient mapping* $\mathcal{G}^\lambda_\phi(x)$ as

$$\mathcal{G}^\lambda_\phi(x) = \frac{1}{\lambda}(x - \text{prox}_{\lambda h}(x - \lambda\nabla f(x))). \tag{13}$$

## 2 Primal Algorithm: Restarted ACG

In this section, we consider the CSCO problem (1) under the following standard assumptions.

**Assumption 1.** Problem (1) satisfies the following:

(a) $f$ is proper closed, $\mu_f$-strongly convex and $L_f$-smooth on $\mathbb{R}^n$,

(b) the smoothness parameter $L_f$ and strong convexity parameter $\mu_f$ satisfy $L_f \geq 2\mu_f \geq 0$,[1]

(c) $h$ is proper closed convex with a simple proximal mapping.

In Subsection 2.1, we begin by introducing a variant of ACG (Algorithm 1) for solving a regularized version of problem (1), and we establish its convergence rate bound as an inner solver. Building on this result, Subsection 2.2 proposes a Restarted ACG method that repeatedly invokes Algorithm 1 to solve a sequence of proximal subproblems of (1). We analyze the outer iteration complexity of this restarted method, and by combining the inner complexity of Algorithm 1, we derive the overall complexity of the Restarted ACG method.

### 2.1 Overview of an ACG variant

Throughout this work, we will utilize ACG as a subroutine to solve regularized subproblems. The generic regularized subproblem we consider is of the form

$$\min\{\psi(x) := g(x) + h(x) : x \in \mathbb{R}^n\}, \tag{14}$$

where $g$ is $\mu$-strongly convex and $(L+\mu)$-smooth, and $h$ is a convex and possibly nonsmooth function with a simple proximal mapping, satisfying $\operatorname{dom} h \subset \operatorname{dom} g$. We describe an ACG variant tailored to (14) and present some basic results regarding the ACG variant.

**Algorithm 1** Accelerated Composite Gradient

**Initialize:** given initial point $x_0 \in \operatorname{dom}\psi$, $L \geq 0$, and $\mu \geq 0$, set $A_0 = 0$, $\tau_0 = 1$, and $y_0 = x_0$.

**for** $j = 0, 1, \cdots$ **do**

**1.** Compute

$$a_j = \frac{\tau_j + \sqrt{\tau_j^2 + 8\tau_j A_j L}}{4L}, \quad A_{j+1} = A_j + a_j, \quad \tau_{j+1} = \tau_j + \mu a_j, \tag{15}$$

$$\tilde{x}_j = \frac{A_j}{A_{j+1}} y_j + \frac{a_j}{A_{j+1}} x_j. \tag{16}$$

**2.** Compute

$$\tilde{y}_{j+1} = \operatorname*{argmin}_{u \in \mathbb{R}^n} \left\{ \ell_g(u; \tilde{x}_j) + h(u) + \frac{2L+\mu}{2}\|u - \tilde{x}_j\|^2 \right\}, \tag{17}$$

$$y_{j+1} = \operatorname{argmin} \{\psi(y_j), \psi(\tilde{y}_{j+1})\}, \tag{18}$$

$$x_{j+1} = \frac{(2L+\mu)a_j\tilde{y}_{j+1} - \frac{2A_j a_j L}{A_{j+1}} y_j}{A_{j+1}\mu + 1}. \tag{19}$$

**end for**

[1] This assumption can be made without loss of generality, since the definition of $L_f$-smoothness in (12) implies $f$ is also $2L_f$-smooth. In view of Theorem 2.3, this only incurs a constant $\sqrt{2}$ factor increase in the iteration complexity of Algorithm 2.

The following convergence rates are standard for first-order accelerated methods. However, we provide a self-contained analysis of Algorithm 1 based on the FLOrA framework in Appendix D.1 for completeness.

**Lemma 2.1.** *Define $R_0 = \min\{\|x - x_0\| : x \in X_*\}$, where $X_*$ is the set of optimal solutions to (14). Then, for all $j \geq 1$,*

$$\psi(y_j) - \psi(x_*) \leq \frac{R_0^2}{2A_j}, \tag{20}$$

$$\|\mathcal{G}_\psi^{(2L+\mu)^{-1}}(\tilde{x}_{j-1})\| \leq \frac{(2L+\mu)R_0}{\sqrt{LA_j}}. \tag{21}$$

The following lemma develops technical bounds in terms of the relative quantity $\|y_j - x_0\|$. These bounds will be critical in analyzing the Restarted ACG algorithm proposed in Subsection 2.2. The proof is deferred to Appendix D.1.

**Lemma 2.2.** *For every $j \geq 1$, define*

$$\Gamma_j(\cdot) := \ell_g(\cdot; \tilde{x}_j) + h(\cdot) + \frac{2L+\mu}{2}\|u - \tilde{x}_j\|^2,$$

$$\theta_{j+1}(x) := \Gamma_j(\tilde{y}_{j+1}) - L\|\tilde{y}_{j+1} - \tilde{x}_j\|^2 + \langle u_{j+1}, x - \tilde{y}_{j+1}\rangle + \frac{\mu}{2}\|x - \tilde{y}_{j+1}\|^2, \tag{22}$$

$$\Theta_{j+1}(x) := \frac{A_j\Theta_j(x) + a_j\theta_{j+1}(x)}{A_{j+1}}, \tag{23}$$

$$\hat{x}_j := \operatorname*{argmin}_{u\in\mathbb{R}^n}\{\Theta_j(u)\}, \qquad s_j := \frac{x_0 - x_j}{A_j} \in \partial\Theta_j(x_j), \tag{24}$$

*where $\Theta_0(\cdot) = 0$. Assuming that $A_j \geq 3/\mu$, then the following statements hold for every $j \geq 1$:*

a)

$$\psi(y_j) - \Theta_j(\hat{x}_j) \leq \frac{\mu}{\mu A_j - 2}\|y_j - x_0\|^2; \tag{25}$$

b)

$$\|s_j\| \leq \frac{3\|y_j - x_0\|}{2A_j}. \tag{26}$$

## 2.2 The Restarted ACG Method

This subsection presents the Restarted ACG method to solve (1). Restarted ACG requires repeatedly invoking Algorithm 1 as a subroutine within a double-loop algorithm. This approach aligns naturally with the IPP framework, which iteratively solves a sequence of proximal subproblems using a recursive subroutine. Within each loop of the IPP framework, Algorithm 1 is employed to solve a certain proximal subproblem, while between successive loops, an acceleration scheme is applied. Consequently, the proposed Restarted ACG method (i.e., Algorithm 2) can be described as "doubly accelerated."

**Algorithm 2** Restarted ACG

**Initialize:** given initial point $w_0 \in \operatorname{dom} h$, $\sigma \in (0,1)$, $L_f \geq 0$, $\mu_f \geq 0$, and $\lambda > 0$, set $B_0 = 0$, $\tau_0 = 1$, and $v_0 = w_0$.

**for** $k = 0, 1, \cdots$ **do**

**1.** Compute

$$b_k = \frac{\tau_k\lambda + \sqrt{\tau_k^2\lambda^2 + 4\tau_k\lambda B_k}}{2}, \quad B_{k+1} = B_k + b_k, \quad \tau_{k+1} = \tau_k + b_k\mu_f,$$

$$\tilde{v}_k = \frac{B_k}{B_{k+1}}w_k + \frac{b_k}{B_{k+1}}v_k. \tag{27}$$

**2.** Call Algorithm 1 with

$$x_0 = \tilde{v}_k, \quad \psi(\cdot) = g(\cdot) + h(\cdot), \quad g(\cdot) = f(\cdot) + \frac{1}{2\lambda}\|\cdot - \tilde{v}_k\|^2, \quad \mu = \mu_f + \frac{1}{\lambda}, \quad L = L_f - \mu_f \tag{28}$$

and perform $j$ iterations until

$$\|\lambda s_j\|^2 + 2\lambda[\psi(y_j) - \Theta_j(x_j)] \leq \sigma\|y_j - x_0\|^2, \tag{29}$$

where $y_j$ and $x_j$ are the ACG iterates defined in (18) and (19), respectively, and $\Theta_j$ and $s_j$ are defined in (23) and (24), respectively.

**3.** Choose $w_{k+1} \in \operatorname{Argmin}\{\phi(u) : u \in \{w_k, y_j\}\}$ and compute

$$v_{k+1} = \frac{1}{\tau_{k+1}}\left(\tau_k v_k + b_k\mu_f x_j - b_k\frac{A_j + \lambda}{\lambda}s_j\right), \tag{30}$$

where $A_j$ is the ACG scalar as in (15).

**end for**

From the "inner loop" perspective, Algorithm 2 keeps performing ACG iterations to solve the proximal subproblem (14) with specification as in (28) until (29) is satisfied, and then restarts ACG with the initialization as in (28). From the "outer loop" perspective, Algorithm 2 is an instance of the FLOrA framework for solving (1) with ACG as its subroutine to implement Step **2** of Algorithm 6, as we will show in Subsection 5.1.

The next result combines the "outer" and "inner" complexities (see Propositions 5.2 and 5.3, respectively) to obtain the total iteration-complexity of Algorithm 2.

**Theorem 2.3.** *For given $\varepsilon > 0$, the following statements hold:*

(a) *if $\mu_f = 0$ and $1/L_f \leq \lambda \leq R_0^2/\varepsilon$, then the total iteration-complexity of Algorithm 2 to find an $\varepsilon$-solution is $\tilde{\mathcal{O}}\left(R_0\sqrt{L_f/\varepsilon}\right)$;*

(b) *if $\mu_f > 0$ and $1/(L_f - \mu_f) \leq \lambda \leq \min\{1/\mu_f, R_0^2/\varepsilon\}$, then the total iteration-complexity of Algorithm 2 to find an $\varepsilon$-solution is $\tilde{\mathcal{O}}(\min\{\sqrt{L_f/\mu_f}, R_0\sqrt{L_f/\varepsilon}\})$.*

If $\lambda$ is taken to be $1/(L_f - \mu_f)$, we can show that the number of ACG iterations on each call to Algorithm 1 is $\mathcal{O}(1)$ (see (71) below). If $\lambda$ is taken sufficiently small, then each call will only perform a single ACG iteration, effectively reducing Restarted ACG (i.e., Algorithm 2) to standard ACG (i.e., Algorithm 1).

# 3 Dual Algorithm: Augmented Lagrangian

In this section we consider the LC-CSCO problem (2) under the following standard assumptions.

**Assumption 2.** Problem (2) satisfies the following:

(a) $f$ is proper closed convex and $L_f$-smooth on $\mathbb{R}^n$,

(b) $h$ is proper closed convex with a simple proximal mapping,

(c) Slater's condition (i.e., (11)) is satisfied,

(d) $\operatorname{dom} h$ is bounded with diameter $D \geq 1$.

The exact ALM (see (6) and (7)) was originally developed from the primal perspective [15], with the quadratic term $\rho\|Ax - b\|/2$ motivated by explicit penalty methods. However, as noted by Rockafellar [42], the ALM can be reformulated as a proximal point method in the dual,

$$\lambda_{k+1} = \operatorname*{argmax}_{\lambda \in \mathbb{R}^m} \left\{ d(\lambda) - \frac{1}{2\rho}\|\lambda - \lambda_k\|^2 \right\}, \tag{31}$$

where $\rho > 0$ is now the proximal stepsize. As mentioned in Section 1, however, solving (31) (i.e., (6)) is typically intractable. Accordingly, practitioners instead adopt the I-ALM with the inexact primal step (9). Since ALM is equivalent to the proximal point method, it is only natural to suppose that I-ALM is equivalent to the IPP iteration

$$\lambda_{k+1} \approx \operatorname*{argmax}_{\lambda \in \mathbb{R}^m} \left\{ d(\lambda) - \frac{1}{2\rho}\|\lambda - \lambda_k\|^2 \right\},$$

for some suitable definition of "inexactness". Letting $\hat{\lambda}_k$ be the exact minimizer to the dual proximal problem (31), Rockafellar [42, Proposition 6] proved that

$$\frac{1}{2\rho}\|\lambda^{k+1} - \hat{\lambda}_k\|^2 \leq \mathcal{L}_\rho(x_{k+1}, \lambda_k) - \min_{x \in \mathbb{R}^n} \mathcal{L}_\rho(x, \lambda_k).$$

Thus, if we can ensure $\mathcal{L}_\rho(x_{k+1}, \lambda_k) - \min_{x \in \mathbb{R}^n} \mathcal{L}_\rho(x, \lambda_k) \leq \varepsilon_k$ for some summable sequence $\{\varepsilon_k\}_{k=0}^\infty$, then we can show (see [42, Theorem 4]) that $\lim_{k\to\infty} \lambda_k = \lambda_*$ for some $\lambda_* \in \{\lambda : d(\lambda) = \max_{\nu \in \mathbb{R}^m} d(\nu)\}$. This "absolute error" IPP perspective has persisted in several recent analyses of the I-ALM [25, 53].

Instead of a traditional absolute error framework, we use the LOrA and FLOrA frameworks of Section 4 to provide an IPP perspective on the I-ALM, enabling us to mix relative and absolute error criteria. Subsection 3.1 provides near-optimal iteration-complexity bounds for a baseline I-ALM (Algorithm 3), improving on the non-ergodic complexity bounds from [18, 25]. Utilizing the FLOrA framework, Subsection 3.2 then proposes an accelerated ALM variant, I-FALM (Algorithm 4).

## 3.1 Inexact Augmented Lagrangian Method

In this subsection we prove near-optimal, non-ergodic $\tilde{\mathcal{O}}(\varepsilon^{-1})$ complexity for the I-ALM (Algorithm 3) with constant penalty term $\rho > 0$. For convenience, we denote the smooth part of the augmented Lagrangian (8) as

$$\Psi_\lambda(x) := f(x) + \langle \lambda, Ax - b \rangle + \frac{\rho}{2}\|Ax - b\|^2,$$

with smoothness constant $M_\rho$

$$M_\rho := L_f + \rho\|A\|^2. \tag{32}$$

The termination condition for the inexact iteration (9) is typically stated in terms of an $\varepsilon_k$-small objective gap, i.e., $\mathcal{L}_\rho(x_{k+1}, \lambda_k) - \min_{x\in\mathbb{R}^n} \mathcal{L}_\rho(x, \lambda_k) \le \varepsilon_k$, for some specified tolerance $\varepsilon_k > 0$. In most cases, however, the exact objective gap is not computable. Supposing that $x_{k+1}$ is computed from a proximal mapping with stepsize $\eta < 1/L_f$, i.e., $x_{k+1} = \mathrm{prox}_{\eta h}(\tilde{x}_k - \eta\nabla f(\tilde{x}_k))$, we can use an alternative termination condition based on the gradient mapping $\mathcal{G}^\eta_{\mathcal{L}_\rho(\cdot,\lambda_k)}(\tilde{x}_k)$, defined in (13). Applying Lemma B.1(a) with $\tilde{x} = \tilde{x}_k$ and $x^+ = x_{k+1}$, the condition $\|\mathcal{G}^\eta_{\mathcal{L}_\rho(\cdot,\lambda_k)}(\tilde{x}_k)\| \le \varepsilon_k/D$ implies $\mathcal{L}_\rho(x_{k+1}, \lambda_k) - \min_{x\in\mathbb{R}^n} \mathcal{L}_\rho(x, \lambda_k) \le \varepsilon_k$. Unlike the primal gap, the gradient mapping norm is an explicit and efficiently computable quantity: providing a practical inner termination condition.

However, if the objective $\mathcal{L}_\rho(\cdot, \lambda_k)$ is merely convex, then we can show that Algorithm 1 requires $\mathcal{O}(\varepsilon_k^{-2/3})$ iterations to guarantee an $\varepsilon_k$-small gradient mapping [35], worse than the $\mathcal{O}(\varepsilon_k^{-1/2})$ complexity needed for an $\varepsilon_k$-small primal gap. To improve the complexity of the inner call, we can instead optimize the objective

$$\min_{x\in\mathbb{R}^n} \left\{ \mathcal{L}_\rho(x, \lambda_k) + \frac{\varepsilon_k}{4D^2}\|x - x_k\|^2 \right\},$$

which, as shown in Proposition 5.4 below, guarantees $\tilde{\mathcal{O}}(D/\sqrt{\varepsilon_k})$ complexity for each inner iteration. This "perturbation" trick is common for improving the complexity of finding a gradient [36, Subsection 2.2.2] or gradient mapping [35, Subsection 5.2] with $\varepsilon_k$-small norm. A gradient mapping termination criterion also removes the need for post-processing routines (e.g., [18]) and improves theoretical guarantees with fixed $\rho$, as we discuss further in remarks following Theorem 3.1.

**Algorithm 3** Inexact Augmented Lagrangian Method

**Initialize:** given initial point $x_0 \in \mathrm{dom}\, h$, $\rho > 0$, $\varepsilon_0 > 0$, $\alpha \in (0, 1)$, $\varepsilon > 0$, set $\lambda_0 = 0$, and choose $\sigma \in (0, 1)$ such that $2\sigma\rho \le D/\varepsilon$.

**for** $k = 0, 1, \cdots$ **do**

**1.** Set $\varepsilon_k = (\varepsilon_0\alpha^k + \sigma\rho\varepsilon^2)/2$ and call Algorithm 1 with

$$\begin{gathered} x_0 = x_k, \quad \psi(\cdot) = \mathcal{L}_\rho(\cdot, \lambda_k) + \frac{\varepsilon_k}{8D^2}\|\cdot - x_k\|^2, \quad g(\cdot) = \Psi_{\lambda_k}(\cdot) + \frac{\varepsilon_k}{8D^2}\|\cdot - x_k\|^2, \\ L = M_\rho, \quad \mu = \frac{\varepsilon_k}{4D^2}, \end{gathered} \tag{33}$$

to find a $\tilde{x}_k$ satisfying $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\mathcal{L}_\rho(\cdot,\lambda_k)}(\tilde{x}_k)\| \le \varepsilon_k/(2D)$ and set

$$x_{k+1} = \tilde{x}_k - (2L + \mu)^{-1}\mathcal{G}^{(2L+\mu)^{-1}}_{\mathcal{L}_\rho(\cdot,\lambda_k)}(\tilde{x}_k).$$

**2.** Compute

$$\lambda_{k+1} = \lambda_k + \rho(Ax_{k+1} - b). \tag{34}$$

**3.** If $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\mathcal{L}_\rho(\cdot,\lambda_k)}(\tilde{x}_k)\| \le \varepsilon/2$ and $\|Ax_{k+1} - b\| \le \varepsilon$, then **return** $(x_{k+1}, \lambda_{k+1})$.

**end for**

To analyze Algorithm 3, we separately consider the "inner" and "outer" perspectives. For the inner, we apply the known iteration-complexity guarantees of ACG to achieve the termination

condition in Step **2** (see Proposition 5.4 below). For the outer, we first prove I-ALM as an instance of the LOrA framework (see Section 4.1 below), and then apply the sub-optimality guarantee of LOrA to obtain the outer iteration-complexity. Combining the two perspectives yields the following iteration-complexity bound, whose proof is deferred to Subsection 5.2.

**Theorem 3.1.** *Given $\varepsilon > 0$, we choose $\varepsilon_0 = \varepsilon$, $\sigma = 1/2$, and $\rho = \varepsilon^{-1}$. Then, Algorithm 3 finds an $\varepsilon$-primal-dual solution to* (2) *in*

$$\tilde{\mathcal{O}}\left((1 + R_\Lambda^2)\left(1 + D\left(\frac{\sqrt{L_f}}{\sqrt{\varepsilon}} + \frac{\|A\|}{\varepsilon}\right)\right)\right) \tag{35}$$

*ACG iterations, where $R_\Lambda = \|\lambda_* - \lambda_0\| = \min\{\|\lambda - \lambda_0\| : \lambda \in \Lambda_*\}$, where $\Lambda_*$ is the set of maximizers to the dual problem* (3).

**Remark.** Lan and Monteiro [18] analyzed I-ALM with a static $\rho$ similar to Algorithm 3, obtaining an iteration-complexity of $\mathcal{O}(\varepsilon^{-7/4})$. While there are a number of differences between Lan and Monteiro's approach and ours, the $\varepsilon$-complexity disparity can be attributed primarily to the method of ensuring $\varepsilon$-stationarity. Lan and Monteiro used a "refinement" final call to Algorithm 1 with $\mathcal{O}(M_\rho^{-1}\varepsilon^2)$ accuracy, thereby requiring $\mathcal{O}(M_\rho/\varepsilon) = \mathcal{O}((L_f + \rho\|A\|^2)/\varepsilon)$ iterations. Setting $\rho = \varepsilon^{-1}$ would result in the "refinement" phase taking $\mathcal{O}(\varepsilon^{-2})$ iterations, requiring the authors to trade off "main loop" and "refinement" complexity. In contrast, our usage of gradient mapping norms to provide stationarity guarantees does not require a postprocessing stage, and each inner call takes $\mathcal{O}((\sqrt{L_f} + \sqrt{\rho}\|A\|)/(\varepsilon\sqrt{\rho}))$ iterations (see (74) below), enabling us to take $\rho = \varepsilon^{-1}$ without adding superfluous $\varepsilon$-dependence.

Using Lemma B.3 from Appendix B, we can translate the $\varepsilon$-solution complexity in Theorem 3.1 into the complexity to find an $\varepsilon$-primal solution in the sense of (5). The proof is deferred to Appendix E.

**Corollary 3.2.** *Under the conditions and parameter choices of Theorem 3.1, Algorithm 3 finds an $\varepsilon$-primal solution to* (2) *in*

$$\tilde{\mathcal{O}}\left((1 + R_\Lambda^2)\left(1 + D\left(\frac{\sqrt{(R_\Lambda + D)L_f}}{\sqrt{\varepsilon}} + \frac{(R_\Lambda + D)\|A\|}{\varepsilon}\right)\right)\right) \tag{36}$$

*ACG iterations, where $R_\Lambda = \|\lambda_* - \lambda_0\| = \min\{\|\lambda - \lambda_0\| : \lambda \in \Lambda_*\}$, and $\Lambda_*$ is the set of maximizers to the dual problem* (3).

**Remark.** Comparing to the lower bounds for primal convergence established in [39, Theorem 3.1], the complexity of Corollary 3.2 is optimal (up to logarithmic terms) in $\varepsilon$, $L_f$, and $\|A\|$. However, it is suboptimal in $R_\Lambda$ ($\mathcal{O}(R_\Lambda^3)$ vs $\mathcal{O}(R_\Lambda)$) and $D$ ($\mathcal{O}(D^2)$ vs $\mathcal{O}(D)$). The discrepancy may be due to our choice of optimality measure in (4). Lu and Zhou [26] also obtained optimal complexity in terms of $L_f$, $\|A\|$, and $\varepsilon$, but similarly incurred additional $R_\Lambda$ and $D$ dependence when converting to a primal gap bound.

### 3.2 Inexact Fast Augmented Lagrangian Method

While Algorithm 3 is near-optimal, numerical experiments in Section 6 below show that it is often outperformed by more advanced methods such as the linearized proximal ALM (LPALM) [21, 38], particularly when $\rho = \varepsilon^{-1}$. In this subsection, we utilize FLOrA (see Subsection 4.2) to accelerate the outer loop, accelerating dual maximization and leading to a more performant algorithm.

As discussed in the last subsection, accelerated methods typically require $\mathcal{O}(\varepsilon^{-2/3})$ iterations to guarantee an $\varepsilon$-small gradient mapping for a merely convex objective. However, adding a small, strongly convex perturbation improves the iteration complexity to $\mathcal{O}(\varepsilon^{-1/2})$. Since the criterion (4) can be interpreted as finding $\varepsilon$-small primal/dual subgradients, we add strongly convex (concave) perturbations to the primal (dual) problems to improve the iteration-complexity. First, we define the perturbed primal problem

$$\tilde{\phi}_* := \min_{x\in\mathbb{R}^n} \left\{\phi(x) + \frac{\gamma_p}{2}\|x - x_0\|^2 : Ax = b\right\} \tag{37}$$

where $\phi(x)$ is as in (2) and $x_0 \in \operatorname{dom} h$ is an arbitrary point. The associated Lagrangian is then

$$\mathcal{L}^{\gamma_p}(x,\lambda) = \phi(x) + \frac{\gamma_p}{2}\|x - x_0\|^2 + \langle\lambda, Ax - b\rangle,$$

with the augmented form

$$\mathcal{L}_\rho^{\gamma_p}(x,\lambda) = \phi(x) + \frac{\gamma_p}{2}\|x - x_0\|^2 + \langle\lambda, Ax - b\rangle + \frac{\rho}{2}\|Ax - b\|^2.$$

Extending the idea of perturbations to the dual problem, we define the *symmetrically perturbed* (augmented) Lagrangian

$$\tilde{\mathcal{L}}(x,\lambda) := \mathcal{L}^{\gamma_p}(x,\lambda) - \frac{\gamma_d}{2}\|\lambda - \lambda_0\|^2, \qquad \tilde{\mathcal{L}}_\rho(x,\lambda) := \mathcal{L}_\rho^{\gamma_p}(x,\lambda) - \frac{\gamma_d}{2}\|\lambda - \lambda_0\|^2;$$

where $\lambda_0 \in \mathbb{R}^m$ is arbitrary. The associated perturbed dual problem is

$$\tilde{d}(\lambda) = \min_{u\in\mathbb{R}^n} \tilde{\mathcal{L}}(u,\lambda) = \min_{u\in\mathbb{R}^n} \mathcal{L}^{\gamma_p}(u,\lambda) - \frac{\gamma_d}{2}\|\lambda - \lambda_0\|^2, \tag{38}$$

which is $\gamma_d$-strongly concave, hence $-\tilde{d}(\lambda)$ is $\gamma_d$-strongly convex. Accordingly, (38) has a unique maximizer $\tilde{\lambda}_*$ with $R_{\tilde{\Lambda}} := \|\tilde{\lambda}_* - \lambda_0\|$. By the definition of $\tilde{d}$ and the superadditivity of min, we have

$$\begin{aligned}
\tilde{d}(\lambda) &\overset{(38)}{=} \min_{x\in\mathbb{R}^n} \left\{\mathcal{L}(x,\lambda) + \frac{\gamma_p}{2}\|x - x_0\|^2\right\} - \frac{\gamma_d}{2}\|\lambda - \lambda_0\|^2 \\
&\geq \min_{x\in\mathbb{R}^n} \mathcal{L}(x,\lambda) + \overbrace{\min_{x\in\mathbb{R}^n} \frac{\gamma_p}{2}\|x - x_0\|^2}^{=0} - \frac{\gamma_d}{2}\|\lambda - \lambda_0\|^2 \overset{(3)}{=} d(\lambda) - \frac{\gamma_d}{2}\|\lambda - \lambda_0\|^2.
\end{aligned} \tag{39}$$

As before, we define $\Psi_\lambda^{\gamma_p}(\cdot)$ as the smooth, primal part of $\tilde{\mathcal{L}}_\rho(\cdot)$

$$\Psi_\lambda^{\gamma_p}(x) := f(x) + \frac{\gamma_p}{2}\|x - x_0\|^2 + \langle\lambda, Ax - b\rangle + \frac{\rho}{2}\|Ax - b\|^2,$$

which is $\gamma_p$-strongly convex and $(M_\rho + \gamma_p)$-smooth on $\mathbb{R}^n$.

For sufficiently small $\gamma_p$, an approximate solution to the perturbed problem implies an approximate solution to the original target problem. The following result is elementary, and similar lemmas have been used in previous works [18], and we therefore defer its proof to Appendix E.

**Lemma 3.3** (Perturbation Solution)**.** *Set $\gamma_p = \varepsilon/(2D)$ and suppose that for the pair $(x,\lambda)$ there exists $v \in \partial\tilde{\mathcal{L}}(\cdot,\lambda)(x)$ satisfying $\|v\| \leq \varepsilon/2$. Then, there exists $v' \in \partial\mathcal{L}(\cdot,\lambda)(x)$ satisfying $\|v'\| \leq \varepsilon$.*

Primal perturbations have been leveraged in several existing works [18, 26] with strong relations to proximal ALM schemes [28]. Dual perturbation, on the other hand, has been less explored, while it has appeared in previous works to improve the iteration-complexity of the outer ALM loop [40].

However, as far as we are aware, the two ideas have not been used in tandem. As noted in [26], the distance from $\lambda_0$ to the optimum of the perturbed dual (defined as $R_{\tilde{\Lambda}}$) depends implicitly on $\gamma_p^{-1} \propto \varepsilon^{-1}$. Interestingly, adding dual regularization removes this hidden dependence, as we will show in Lemma 5.7 below.

Incorporating the acceleration scheme into the outer loop, as well as the perturbations, we obtain the I-FALM, shown in Algorithm 4.

**Algorithm 4** Inexact Fast Augmented Lagrangian Method

**Initialize:** given initial $x_0 \in \operatorname{dom} h$, $\rho > 0$, $\gamma_d > 0$, $\varepsilon > 0$, and $\varepsilon_0 \geq \varepsilon$, set $B_0 = 0$, $\tau_0 = 1$, $\gamma_p = \varepsilon/(2D)$, and $\lambda_0 = \nu_0 = 0$, and choose $\sigma \in (0,1)$ such that $4\sigma\rho\varepsilon \leq 1$, and $\alpha \geq 0$ satisfying $\alpha < (1 + \sqrt{\gamma_d \rho})^{-2}$.

**for** $k = 0, 1, \cdots$ **do**

**1.** Set $\varepsilon_k = (7\varepsilon_0 \alpha^k + \sigma\rho\varepsilon^2)/8$ and compute

$$b_k = \frac{\rho\tau_k + \sqrt{\rho^2\tau_k^2 + 4\rho\tau_k B_k}}{2}, \quad B_{k+1} = B_k + b_k, \quad \tau_{k+1} = \tau_k + b_k\gamma_d;$$
$$\tilde{\nu}_k = \frac{B_k}{B_{k+1}}\lambda_k + \frac{b_k}{B_{k+1}}\nu_k. \tag{40}$$

**2.** Call Algorithm 1 with

$$x_0 = x_k, \quad \psi(\cdot) = \tilde{\mathcal{L}}_\rho(\cdot, \tilde{\nu}_k) + \frac{\varepsilon_k}{8D^2}\|\cdot - x_k\|^2, \quad g(\cdot) = \Psi_{\tilde{\nu}_k}^{\gamma_p} + \frac{\varepsilon_k}{8D^2}\|\cdot - x_k\|^2,$$
$$L = M_\rho, \quad \mu = \gamma_p + \frac{\varepsilon_k}{4D^2}, \tag{41}$$

to find a point $\tilde{x}_k$ satisfying $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\tilde{\mathcal{L}}_\rho(\cdot,\tilde{\nu}_k)}(\tilde{x}_k)\| \leq \varepsilon_k/(2D)$ and set

$$x_{k+1} = \tilde{x}_k - (2L+\mu)^{-1}\mathcal{G}^{(2L+\mu)^{-1}}_{\tilde{\mathcal{L}}_\rho(\cdot,\tilde{\nu}_k)}(\tilde{x}_k).$$

**3.** Compute

$$\lambda_{k+1} = \tilde{\nu}_k + \rho(Ax_{k+1} - b). \tag{42}$$

**4.** If $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\tilde{\mathcal{L}}_\rho(\cdot,\tilde{\nu}_k)}(\tilde{x}_k)\| \leq \varepsilon/4$ and $\|Ax_{k+1} - b\| \leq \varepsilon$, then return $(x_{k+1}, \lambda_{k+1})$ and **terminate**; otherwise, compute

$$\nu_{k+1} = \frac{1}{\tau_{k+1}}\left(\tau_k\nu_k + b_k\gamma_d\frac{\lambda_{k+1}}{1+\gamma_d\rho} - \frac{b_k}{\rho}\left(\tilde{\nu}_k - \frac{\lambda_{k+1}}{1+\gamma_d\rho}\right)\right), \tag{43}$$

and continue.

**end for**

Unlike Algorithm 3, the $\varepsilon_k/(8D^2)$ addition to $\psi$ in Algorithm 4 is not necessary to obtain $\varepsilon^{-1}$ complexity, as shown in Proposition 5.8 below. However, the increased strong convexity modulus improves empirical performance, particularly in early iterations when $\varepsilon_k \gg \varepsilon$.

Theorem 3.4, whose proof is deferred to Subsection 5.3, states our main complexity results for

Algorithm 4.

**Theorem 3.4.** *Let $\varepsilon > 0$ satisfy $\varepsilon \leq \|A\|^2/L_f$. Choose $\rho = L_f/\|A\|^2$, $\varepsilon_0 = \rho^{-1}$, $\sigma = 1/4$, $\gamma_p = \varepsilon/(2D)$, $\gamma_d = \sigma^{3/2}\varepsilon/(\sqrt{3}\mathcal{R})$, where*

$$\mathcal{R} := \hat{R}_{\tilde{\Lambda}}(1+\sqrt{2\varepsilon_0 C})\left(\frac{2}{\sqrt{1-\sigma}}+1\right), \quad \hat{R}_{\tilde{\Lambda}} := \max\{1, \|\tilde{\lambda}_* - \lambda_0\|\}, \quad C := \sum_{i=0}^{\infty} B_{i+1}\alpha^i < \infty, \tag{44}$$

*with $\tilde{\lambda}_*$ defined as the unique maximizer of (38). Furthermore, assume $\alpha$ satisfies*

$$\alpha \leq \min\left\{\frac{9}{10}(1+\sqrt{\rho\gamma_d})^{-2}, \left(\frac{15D\varepsilon}{28\varepsilon_0}\right)^{\sqrt{\rho\varepsilon/D}}\right\}. \tag{45}$$

*Then, Algorithm 4 finds an $\varepsilon$-primal-dual solution to (2) in*

$$\tilde{\mathcal{O}}\left(1 + \frac{\sqrt{D^2 + D\hat{R}_\Lambda}\|A\|}{\varepsilon} + \frac{\sqrt{D+\hat{R}_\Lambda}\|A\|}{\sqrt{L_f\varepsilon}} + \frac{\sqrt{DL_f}}{\sqrt{\varepsilon}}\right) \tag{46}$$

*total ACG iterations, where*

$$\hat{R}_\Lambda = \max\{1, \|\lambda_* - \lambda_0\|\}, \tag{47}$$

*with $\lambda_* = \operatorname{argmin}\{\|\lambda - \lambda_0\| : \lambda \in \Lambda_*\}$ and $\Lambda_*$ is the set of maximizers to the dual problem (3).*

Combining the previous complexity results with Lemma B.3 in Appendix B, we can state complexity results for obtaining an $\varepsilon_g$-primal solution (see 5). As in the previous subsection, the proof is deferred to Appendix E.

**Corollary 3.5.** *Let $\varepsilon_g > 0$ satisfy $\varepsilon_g \leq 2\|A\|^2(D+\mathcal{R})/L_f$. Then, using the parameter settings of Theorem 3.4 with $\varepsilon = \varepsilon_g/(2(D+\hat{R}_\Lambda))$, Algorithm 4 finds an $\varepsilon_g$-primal solution to (2) in*

$$\tilde{\mathcal{O}}\left(\sqrt{\hat{R}_\Lambda + D}\left(1 + \frac{\sqrt{D}(D+\hat{R}_\Lambda)\|A\|}{\varepsilon_g} + \frac{\sqrt{D+\hat{R}_\Lambda}\|A\|}{\sqrt{L_f\varepsilon_g}} + \frac{\sqrt{DL_f}}{\sqrt{\varepsilon_g}}\right)\right) \tag{48}$$

*ACG iterations, where $\mathcal{R}$ is as in (44) and $\hat{R}_\Lambda$ is as in (47).*

Comparing to the lower bound in [39, Theorem 3.1], Algorithm 4 is therefore optimal up to logarithmic terms in $\varepsilon_g$, $\|A\|$, and $L_f$. Again, however, it is sub-optimal in $\hat{R}_\Lambda$ ($\mathcal{O}(\hat{R}_\Lambda^{3/2})$ vs. $\mathcal{O}(\hat{R}_\Lambda)$) and $D$ ($\mathcal{O}(D^2)$ vs. $\mathcal{O}(D)$). As noted following Corollary 3.2, this is likely due to our method of analysis: reducing from approximate stationarity to a primal gap instead of directly bounding a gap function as in [38].

**Remark.** To prove complexity for the case where $\varepsilon L_f \geq \|A\|^2$, we can simply rescale the objective $\phi(\cdot)$ by a scalar $\chi = \|A\|^2/(\varepsilon L_f) \leq 1$ and find a $\chi\varepsilon$-solution using the settings of Theorem 3.4. Furthermore, we can show by elementary algebra that if $(x_*, \chi\lambda_*)$ is an optimal pair for the rescaled problem, then $(x_*, \lambda_*)$ is an optimal pair for the original problem (2). Since $\|\chi\lambda\| \leq \|\lambda\|$, the distance to $\Lambda_*$ from $\lambda_0 = 0$ does not increase, and we only need to consider the effects on $L_f$ and $\varepsilon$. Therefore, Theorem 3.4 provides complexity bounds for (2) without loss of generality.

**Remark.** Focusing on the regime where $L_f \geq \varepsilon_g$ (true of most problems of interest), Corollary 3.5 implies that, omitting $\hat{R}_\Lambda$ and $D$ dependence, Algorithm 4 has an iteration-complexity of $\tilde{\mathcal{O}}\left(\sqrt{L_f/\varepsilon_g} + \|A\|/\varepsilon_g\right)$. In the case where $2(D+\mathcal{R})\|A\|^2 \leq \varepsilon_g L_f$, following a similar rescaling argument as in the previous remark, we establish an iteration-complexity of $\tilde{\mathcal{O}}\left(\sqrt{L_f/\varepsilon_g} + L_f/\|A\|\right)$.

# 4 Frameworks for Generic Convex Optimization

In this section, we consider the generic optimization problem

$$\Phi_* = \min\{\Phi(x) : x \in \mathbb{R}^n\}, \tag{49}$$

where $\Phi$ is proper, lower semi-continuous, and $\mu$-strongly convex for some $\mu \geq 0$ (with $\mu = 0$ corresponding to the merely convex case). Motivated by IPP frameworks [31, 46], we propose two general schemes for solving (49): a baseline (unaccelerated) framework and an accelerated counterpart in the spirit of accelerated gradient methods. Both frameworks rely on an abstract subroutine that prescribes the accuracy to which each proximal subproblem is solved. Under the assumption that such a subroutine is available, the main results of this section are the sub-optimality guarantees for the two frameworks. These guarantees will be used in Section 5 to establish the iteration-complexity bounds of Restarted ACG, I-ALM, and I-FALM, described in Sections 2 and 3, which are special instances of the frameworks in primal and dual spaces.

## 4.1 Lower Oracle Approximation Framework

This subsection presents the baseline framework, LOrA, given in Algorithm 5 below. We mark LOrA iterates (resp., parameters) with a superscript (resp., subscript) "L" to distinguish from those of specific implementations. For simplicity of presentation and analysis, we assume for this subsection that $\Phi$ is merely convex in (49).

**Algorithm 5** LOrA Framework

**Initialize:** given initial point $x_0^{\mathrm{L}} \in \operatorname{dom}\Phi$, $\sigma_{\mathrm{L}} \in (0,1)$, $\lambda_{\mathrm{L}} > 0$, set $y_0^{\mathrm{L}} = x_0^{\mathrm{L}}$.

**for** $k = 0, 1, \ldots$ **do**

**1.** Choose $\delta_k^{\mathrm{L}} > 0$.

**2.** Find $(y_{k+1}^{\mathrm{L}}, \Gamma_k^{\mathrm{L}}) \in \operatorname{dom}\Phi \times \overline{\operatorname{Conv}}_{1/\lambda_{\mathrm{L}}}(\operatorname{dom}\Phi)$ such that

$$\Gamma_k^{\mathrm{L}}(\cdot) \leq \Phi(\cdot) + \frac{1}{2\lambda_{\mathrm{L}}}\|\cdot - x_k^{\mathrm{L}}\|^2, \tag{50}$$

$$\|\lambda_{\mathrm{L}}\hat{u}_{k+1}^{\mathrm{L}}\|^2 + 2\lambda_{\mathrm{L}}\left[\Phi(y_{k+1}^{\mathrm{L}}) + \frac{1}{2\lambda_{\mathrm{L}}}\|y_{k+1}^{\mathrm{L}} - x_k^{\mathrm{L}}\|^2 - \Gamma_k^{\mathrm{L}}(x_{k+1}^{\mathrm{L}})\right] \leq \sigma_{\mathrm{L}}\|y_{k+1}^{\mathrm{L}} - x_k^{\mathrm{L}}\|^2 + 2\lambda_{\mathrm{L}}\delta_k^{\mathrm{L}}, \tag{51}$$

where for some $\mathcal{A}_k^{\mathrm{L}} \in (0, \infty]$,

$$x_{k+1}^{\mathrm{L}} = \operatorname*{argmin}_{x \in \mathbb{R}^n}\left\{\Gamma_k^{\mathrm{L}}(x) + \frac{1}{2\mathcal{A}_k^{\mathrm{L}}}\|x - x_k^{\mathrm{L}}\|^2\right\}, \quad \hat{u}_{k+1}^{\mathrm{L}} = \frac{x_k^{\mathrm{L}} - x_{k+1}^{\mathrm{L}}}{\mathcal{A}_k^{\mathrm{L}}}. \tag{52}$$

**end for**

LOrA can be understood as an iterative procedure of finding $\{y_{k+1}^{\mathrm{L}}\}$ via certain subroutines satisfying (51), which provides sub-optimality guarantees as shown below in Theorem 4.4. From the perspective of IPP, $y_{k+1}^{\mathrm{L}}$ is obtained by inexactly solving the proximal subproblem $\min_{z \in \mathbb{R}^n}\{\Phi(z) + \|z - x_k^{\mathrm{L}}\|^2/(2\lambda_{\mathrm{L}})\}$, where the solution accuracy is controlled by the sum of a relative error and an absolute error as on the right-hand side of (51). Moreover, $\{x_{k+1}^{\mathrm{L}}\}$ is an auxiliary sequence obtained as in (52) by (approximately) solving the surrogate function $\Gamma_k^{\mathrm{L}}$, which approximates $\Phi + \|\cdot - x_k^{\mathrm{L}}\|^2/(2\lambda_{\mathrm{L}})$ from below (see (50)).

The following result formalizes the connection to IPP, that is, $y_{k+1}^{\mathrm{L}}$ is an approximate solution to the proximal subproblem. For brevity of the main text, we defer the proof to Appendix C.1.

**Proposition 4.1.** *Let $\hat{x}_*^{\mathrm{L}}$ be the minimizer of the proximal subproblem at iteration $k$,*

$$\hat{x}_*^{\mathrm{L}} = \operatorname*{argmin}_{z\in\mathbb{R}^n} \left\{ \Phi(z) + \frac{1}{2\lambda_{\mathrm{L}}}\|z - x_k^{\mathrm{L}}\|^2 \right\}.$$

*Then, $y_{k+1}^{\mathrm{L}}$ obtained by the LOrA framework satisfies*

$$\Phi(y_{k+1}^{\mathrm{L}}) + \frac{1}{2\lambda_{\mathrm{L}}}\|y_{k+1}^{\mathrm{L}} - x_k^{\mathrm{L}}\|^2 - \Phi(\hat{x}_*^{\mathrm{L}}) - \frac{1}{2\lambda_{\mathrm{L}}}\|\hat{x}_*^{\mathrm{L}} - x_k^{\mathrm{L}}\|^2 \le \frac{\sigma_{\mathrm{L}}}{2\lambda_{\mathrm{L}}}\|y_{k+1}^{\mathrm{L}} - x_k^{\mathrm{L}}\|^2 + \delta_k^{\mathrm{L}}.$$

LOrA is a generic framework for convex optimization that includes many first-order methods for solving smooth and nonsmooth problems as instances. In Subsection 5.2, we will show that I-ALM is an instance of LOrA. Here we provide two other concrete instances of the LOrA framework: the proximal gradient method and the modern proximal bundle (MPB) method [23, 24].

**Example 4.2** (Proximal Gradient Method)**.** Consider solving problem (49) with $\Phi(\cdot) = f(\cdot) + h(\cdot)$ where $f$ is convex and $L_f$-smooth and $h$ is convex and simple, and choose a stepsize $\eta \le 1/L_f$, the proximal gradient method is

$$x_{k+1} = \mathrm{prox}_{\eta h}(x_k - \eta\nabla f(x_k)). \tag{53}$$

It is straightforward to verify that PGM is an instance of LOrA with the correspondence

$$\begin{gathered}\Phi(\cdot) = f(\cdot) + h(\cdot), \quad \Gamma_k^{\mathrm{L}}(\cdot) = \ell_f(\cdot; x_k) + h(\cdot) + \frac{1}{2\eta}\|\cdot - x_k\|^2, \quad \lambda_{\mathrm{L}} = \eta, \quad \sigma_{\mathrm{L}} = \eta L_f; \\ \mathcal{A}_k^{\mathrm{L}} = \infty, \quad \delta_k^{\mathrm{L}} = 0, \quad y_{k+1}^{\mathrm{L}} = x_{k+1}^{\mathrm{L}} = x_{k+1}, \quad \hat{u}_{k+1}^{\mathrm{L}} = 0.\end{gathered} \tag{54}$$

We can easily verify (50)-(52). First, the inequality (50) follows trivially by the convexity of $f$. Second, it is easy to verify that (53) indicates that

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{ \ell_f(x; x_k) + h(x) + \frac{1}{2\eta}\|x - x_k\|^2 \right\},$$

which in view of the choices of $\Gamma_k^{\mathrm{L}}$, $\mathcal{A}_k^{\mathrm{L}}$, and $x_k^{\mathrm{L}}$ in (54) implies that the first relation in (52) holds. Moreover, the second relation in (52) also simply follows from (54). Finally, we only need to show (51). Using the choices of $y_{k+1}^{\mathrm{L}}$, $x_{k+1}^{\mathrm{L}}$, and $\Gamma_k^{\mathrm{L}}$ in (54), we have

$$\begin{aligned}&2\lambda_{\mathrm{L}}\left[\Phi(y_{k+1}^{\mathrm{L}}) + \frac{1}{2\lambda_{\mathrm{L}}}\|y_{k+1}^{\mathrm{L}} - x_k^{\mathrm{L}}\|^2 - \Gamma_k^{\mathrm{L}}(x_{k+1}^{\mathrm{L}})\right] \\ \overset{(54)}{=}\ &2\eta[f(x_{k+1}) - \ell_f(x_{k+1}; x_k)] \le \eta L_f\|x_{k+1} - x_k\|^2 \overset{(54)}{=} \sigma_{\mathrm{L}}\|y_{k+1}^{\mathrm{L}} - x_k^{\mathrm{L}}\|^2.\end{aligned}$$

where the inequality follows from the $L_f$-smoothness of $f$ and the final identity follows from the choice of $\sigma_{\mathrm{L}}$ in (54).

**Example 4.3** (MPB Method)**.** Consider the composite nonsmooth convex optimization problem $\min_x\{\phi(x) := f(x) + h(x)\}$ where $f$ is convex and Lipschitz continuous and $h$ is convex and simple. One method to solve such problem is the MPB method [23, 24]. A key distinction of MPB from

classical proximal bundle methods [19, 20, 29, 51] lies in its incorporation of the IPP framework. MPB approximately solves a sequence of proximal subproblem of the form

$$\min_{u\in\mathbb{R}^n}\left\{\psi(u):=\phi(u)+\frac{1}{2\lambda}\left\|u-x_k^{\mathrm{L}}\right\|^2\right\}. \tag{55}$$

Letting $x_0=x_k^{\mathrm{L}}$ be the initial point of the subroutine for solving (55), MPB iteratively solves

$$x_j=\operatorname*{argmin}_{u\in\mathbb{R}^n}\left\{\Gamma_j(u)+h(u)+\frac{1}{2\lambda}\|u-x_0\|^2\right\}, \tag{56}$$

where $\Gamma_j$ is a bundle model underneath $f$. Details about various models and a unifying framework underlying them are discussed in [24]. MPB keeps refining $\Gamma_j$ and solving $x_j$ through (56), until a criterion $t_j=\psi(\tilde{x}_j)-m_j\le\delta$ is met, where

$$m_j=\Gamma_j(x_j)+h(x_j)+\frac{1}{2\lambda}\|x_j-x_0\|^2,\quad \tilde{x}_j\in\operatorname{Argmin}\{\psi(u):u\in\{x_0,x_1,\dots,x_j\}\}. \tag{57}$$

As explained in [22], the criterion $t_j\le\delta$ indicates that a primal-dual solution to (55) with primal-dual gap bounded by $\delta$ is obtained. It also implies that $\tilde{x}_j$ is a $\delta$-solution to (55) (see also [23]). Once the condition $t_j\le\delta$ is met, MPB updates the prox center to $x_{k+1}^{\mathrm{L}}=x_j$, resets the bundle model $\Gamma_j$ from scratch, and proceeds to solve (55) with $x_k^{\mathrm{L}}$ replaced by $x_{k+1}^{\mathrm{L}}$. We refer to iterations where the prox center $x_k^{\mathrm{L}}$ is updated (and hence $t_j\le\delta$) as serious steps. Otherwise, a step is referred to as a null step.

Let $j_k$ be the iteration corresponding to serious step $k$. We will show that for all serious steps $k\ge0$, MPB is an instance of the LOrA framework with the correspondence

$$\begin{gathered}\Phi(\cdot)=\phi(\cdot),\quad \Gamma_k^{\mathrm{L}}(\cdot)=\Gamma_{j_k}(\cdot)+h(\cdot)+\frac{1}{2\lambda}\|\cdot-x_k^{\mathrm{L}}\|^2,\quad \lambda_{\mathrm{L}}=\lambda,\quad \sigma_{\mathrm{L}}=0;\\ \mathcal{A}_k^{\mathrm{L}}=\infty,\quad \delta_k^{\mathrm{L}}=\delta,\quad y_{k+1}^{\mathrm{L}}=\tilde{x}_{j_k},\quad x_{k+1}^{\mathrm{L}}=x_{j_k},\quad \hat{u}_{k+1}^{\mathrm{L}}=0.\end{gathered} \tag{58}$$

Inequality (50) follows from the fact that $\Gamma_{j_k}\le f$. The first relation in (52) follows from (56) with $j=j_k$ and the correspondence (58). The second relation in (52) trivially follows from (58). Finally, in view of (58), condition (51) is exactly the serious/null criterion $t_{j_k}\le\delta$, i.e.,

$$\Phi(y_{k+1}^{\mathrm{L}})+\frac{1}{2\lambda_{\mathrm{L}}}\|y_{k+1}^{\mathrm{L}}-x_k^{\mathrm{L}}\|^2-\Gamma_k^{\mathrm{L}}(x_{k+1}^{\mathrm{L}})\overset{(55),(57)}{=}\psi(\tilde{x}_{j_k})-m_{j_k}=t_{j_k}\le\delta.$$

The following theorem presents two sub-optimality guarantees of LOrA. Its proof is deferred to Appendix C.1.

**Theorem 4.4.** *Let $X_*$ be the set of optimal solutions to (49). Define $R_0^{\mathrm{L}}:=\|x_0^{\mathrm{L}}-x_*\|=\min\{\|x_0^{\mathrm{L}}-x\|:x\in X_*\}$ and $\bar{\delta}_k^{\mathrm{L}}:=k^{-1}\sum_{i=0}^{k-1}\delta_i^{\mathrm{L}}$. Suppose that $\mathcal{A}_k^{\mathrm{L}}=\infty$ at all iterations. Then, for every $k\ge1$, we have*

$$\min_{1\le i\le k}\|y_i^{\mathrm{L}}-x_{i-1}^{\mathrm{L}}\|\le\frac{R_0^{\mathrm{L}}}{\sqrt{1-\sigma_{\mathrm{L}}}\sqrt{k}}+\sqrt{\frac{2\lambda_{\mathrm{L}}\bar{\delta}_k^{\mathrm{L}}}{1-\sigma_{\mathrm{L}}}}. \tag{59}$$

*Moreover,*

$$\min_{1\le i\le k}\Phi(y_i^{\mathrm{L}})-\Phi(x_*)\le\frac{\left(R_0^{\mathrm{L}}\right)^2}{2\lambda_{\mathrm{L}}k}+\bar{\delta}_k^{\mathrm{L}}. \tag{60}$$

### 4.2 Fast Lower Oracle Approximation Framework

This subsection presents the FLOrA framework for (strongly) convex minimization. We mark FLOrA iterates (resp., parameters) with a superscript (resp., subscript) "F" to distinguish from those of specific implementations. Incorporating a Nesterov-type acceleration scheme into LOrA, FLOrA achieves better sub-optimality guarantees, which are comparable to other accelerated IPP frameworks [27, 32].

**Algorithm 6** FLOrA Framework

**Initialize:** given initial point $x_0^{\mathrm{F}} \in \operatorname{dom}\Phi$, $\mu_{\mathrm{F}} \geq 0$, $\sigma_{\mathrm{F}} \in (0,1]$, $\lambda_{\mathrm{F}} > 0$, $\tau_0 = 1$, $\delta_0^{\mathrm{F}} \geq 0$, set $B_0 = 0$ and $y_0^{\mathrm{F}} = x_0^{\mathrm{F}}$ and choose an $\alpha_{\mathrm{F}} \geq 0$ satisfying $\alpha_{\mathrm{F}} < (1+\sqrt{\lambda_{\mathrm{F}}\mu_{\mathrm{F}}})^{-2}$.

**for** $k = 0, 1, \ldots$ **do**

**1.** Set $\delta_k^{\mathrm{F}} = \delta_0^{\mathrm{F}}(\alpha_{\mathrm{F}})^k$ and compute

$$b_k = \frac{\lambda_{\mathrm{F}}\tau_k + \sqrt{\lambda_{\mathrm{F}}^2\tau_k^2 + 4\lambda_{\mathrm{F}}\tau_k B_k}}{2}, \quad B_{k+1} = B_k + b_k, \quad \tau_{k+1} = \tau_k + b_k\mu_{\mathrm{F}}, \tag{61}$$

$$\tilde{x}_k^{\mathrm{F}} = \frac{B_k}{B_{k+1}} y_k^{\mathrm{F}} + \frac{b_k}{B_{k+1}} x_k^{\mathrm{F}}. \tag{62}$$

**2.** Find $(\tilde{y}_{k+1}^{\mathrm{F}}, \Gamma_k^{\mathrm{F}}) \in \operatorname{dom}\Phi \times \underset{\mu_{\mathrm{F}}+\lambda_{\mathrm{F}}^{-1}}{\overline{\mathrm{Conv}}}(\operatorname{dom}\Phi)$ such that

$$\Gamma_k^{\mathrm{F}}(\cdot) \leq \Phi(\cdot) + \frac{1}{2\lambda_{\mathrm{F}}}\|\cdot - \tilde{x}_k^{\mathrm{F}}\|^2, \tag{63}$$

$$\|\lambda_{\mathrm{F}}\hat{u}_{k+1}^{\mathrm{F}}\|^2 + 2\lambda_{\mathrm{F}}\left[\Phi(\tilde{y}_{k+1}^{\mathrm{F}}) + \frac{1}{2\lambda_{\mathrm{F}}}\|\tilde{y}_{k+1}^{\mathrm{F}} - \tilde{x}_k^{\mathrm{F}}\|^2 - \Gamma_k^{\mathrm{F}}(z_{k+1}^{\mathrm{F}})\right] \leq \sigma_{\mathrm{F}}\|\tilde{y}_{k+1}^{\mathrm{F}} - \tilde{x}_k^{\mathrm{F}}\|^2 + 2\lambda_{\mathrm{F}}\delta_k^{\mathrm{F}}, \tag{64}$$

where for some $\mathcal{A}_k^{\mathrm{F}} \in (0, \infty]$,

$$z_{k+1}^{\mathrm{F}} = \underset{v\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\Gamma_k^{\mathrm{F}}(v) + \frac{1}{2\mathcal{A}_k^{\mathrm{F}}}\|v - \tilde{x}_k^{\mathrm{F}}\|^2\right\}, \quad \hat{u}_{k+1}^{\mathrm{F}} = \frac{\tilde{x}_k^{\mathrm{F}} - z_{k+1}^{\mathrm{F}}}{\mathcal{A}_k^{\mathrm{F}}}. \tag{65}$$

**3.** Choose $y_{k+1}^{\mathrm{F}}$ satisfying $\Phi(y_{k+1}^{\mathrm{F}}) \leq \Phi(\tilde{y}_{k+1}^{\mathrm{F}})$ and compute

$$u_{k+1}^{\mathrm{F}} = \hat{u}_{k+1}^{\mathrm{F}} + \frac{\tilde{x}_k^{\mathrm{F}} - z_{k+1}^{\mathrm{F}}}{\lambda_{\mathrm{F}}}, \quad x_{k+1}^{\mathrm{F}} = \frac{1}{\tau_{k+1}}\left(\tau_k x_k^{\mathrm{F}} + b_k\mu_{\mathrm{F}} z_{k+1}^{\mathrm{F}} - b_k u_{k+1}^{\mathrm{F}}\right). \tag{66}$$

**end for**

Similar to LOrA, FLOrA does not specify the subroutine used in Step **2** to find $\tilde{y}_{k+1}^{\mathrm{F}}$ and instead describes the requirement (64) on the subroutine to establish sub-optimality guarantees. In addition to LOrA (which is close to Step **2** in FLOrA), FLOrA employs the necessary computation (i.e., Steps **1** and **3**) for Nesterov's acceleration to enable better guarantees. Hence, FLOrA is considered as a generic framework consisting of accelerated methods as special instances. More specifically, we will show that ACG, Restarted ACG, and I-FALM are instances of FLOrA in Appendix D.1, Subsection 5.1, and Subsection 5.3, respectively.

As an accelerated version of LOrA, FLOrA naturally admits an accelerated IPP interpretation. Prior accelerated IPP frameworks focus on more restricted settings: [32] studies (49) in the purely

convex case, while [27] considers the composite form $\Phi = f + h$ with $f$ being convex and $h$ being strongly convex. In contrast, by including $z_{k+1}^{\mathrm{F}}$ in (66), FLOrA accommodates strong convexity in $\Phi$ directly, without imposing any particular decomposition or structural assumptions on $\Phi$.

The following theorem presents three sub-optimality guarantees of FLOrA. Its proof is deferred to Appendix C.2.

**Theorem 4.5.** *Let $X_*$ be the set of optimal solutions to* (49)*. Define $R_0^{\mathrm{F}} := \|x_0^{\mathrm{F}} - x_*\| = \min\{\|x_0^{\mathrm{F}} - x\| : x \in X_*\}$. Then, for every $k \geq 0$,*

$$\Phi(y_{k+1}^{\mathrm{F}}) - \Phi_* \leq \frac{\left(R_0^{\mathrm{F}}\right)^2}{2B_{k+1}} + \frac{\delta_0^{\mathrm{F}} C_{\mathrm{F}}}{B_{k+1}}, \tag{67}$$

*where $C_{\mathrm{F}} := \sum_{i=0}^{\infty} B_{i+1}(\alpha_{\mathrm{F}})^i < \infty$. Furthermore, if $\sigma_{\mathrm{F}} < 1$, then for every $k \geq 0$, we have*

$$\|\tilde{y}_{k+1}^{\mathrm{F}} - \tilde{x}_k^{\mathrm{F}}\| \leq \frac{\sqrt{\lambda_{\mathrm{F}}} R_0^{\mathrm{F}} + \sqrt{2\lambda_{\mathrm{F}}\delta_0^{\mathrm{F}} C_{\mathrm{F}}}}{\sqrt{(1-\sigma_{\mathrm{F}})B_{k+1}}}, \tag{68}$$

$$\min_{0\leq i\leq k} \|\tilde{y}_{i+1}^{\mathrm{F}} - \tilde{x}_i^{\mathrm{F}}\| \leq \frac{\sqrt{\lambda_{\mathrm{F}}} R_0^{\mathrm{F}} + \sqrt{2\lambda_{\mathrm{F}}\delta_0^{\mathrm{F}} C_{\mathrm{F}}}}{\sqrt{(1-\sigma_{\mathrm{F}})\sum_{i=1}^{k+1} B_i}}. \tag{69}$$

# 5 Proofs of Main Complexity Results

This section is devoted to the complexity analysis of the three algorithms studied in this paper: Restarted ACG, I-ALM, and I-FALM. The three subsections provide proofs of the main results for each method, namely Theorems 2.3, 3.1, and 3.4.

## 5.1 Proof of Theorem 2.3

To prove Theorem 2.3, we will first show that, assuming the call to Algorithm 1 terminates in Step **2**, Algorithm 2 is an instance of the FLOrA framework with only relative error (i.e., $\delta_k^{\mathrm{F}} = 0$ for every $k \geq 0$). Theorem 4.5 will then imply the outer complexity. We then bound the number of "inner" iterations required by Algorithm 1 in Step **2** to satisfy (29). Combining the outer and inner complexities gives Theorem 2.3. For brevity of the main text, we defer the proofs of intermediate results to Appendix D.2.

Let $j$ be the final (inner) iteration of Algorithm 1 when invoked in Step **2**, $y_j$ and $x_j$ be the final ACG iterates as in (18) and (19), respectively, $s_j$ be as defined in (24), $A_j$ be the ACG scalar as in (15), and $\Theta_j$ be the aggregate function defined in (23). Then we will show that Algorithm 2 is an instance of the FLOrA framework with the correspondence

$$\begin{gathered}\Phi(\cdot) = \phi(\cdot), \quad \Gamma_k^{\mathrm{F}}(\cdot) = \Theta_j(\cdot), \quad \mathcal{A}_k^{\mathrm{F}} = A_j, \quad \delta_k^{\mathrm{F}} = \alpha_{\mathrm{F}} = 0, \quad \mu_{\mathrm{F}} = \mu_f, \quad \sigma_{\mathrm{F}} = \sigma, \quad \lambda_{\mathrm{F}} = \lambda; \\ y_k^{\mathrm{F}} = w_k,\ x_k^{\mathrm{F}} = v_k,\ \tilde{x}_k^{\mathrm{F}} = \tilde{v}_k,\ \tilde{y}_{k+1}^{\mathrm{F}} = y_j,\ z_{k+1}^{\mathrm{F}} = x_j,\ u_{k+1}^{\mathrm{F}} = \frac{A_j + \lambda}{\lambda} s_j,\ \hat{u}_{k+1}^{\mathrm{F}} = s_j.\end{gathered} \tag{70}$$

**Lemma 5.1.** *Assume for all $k \geq 0$, the call to Algorithm 1 in Step **2** terminates. Then, with the correspondence* (70)*, Algorithm 2 is an instance of the FLOrA framework.*

Since Algorithm 2 is an instance of FLOrA, the following "outer" sub-optimality guarantee holds by Lemma C.2(c) in Appendix C.2 and Theorem 4.5 (see (67) with $\delta_0^{\mathrm{F}} = 0$).

**Proposition 5.2.** *For every $k \geq 1$, the function value gap $\phi(w_k) - \phi_*$ satisfies*

$$\phi(w_k) - \phi_* \leq \min\left\{\frac{2R_0^2}{\lambda k^2}, \quad \frac{R_0^2}{2\lambda}\left(1 + \frac{\sqrt{\lambda\mu_f}}{2}\right)^{-2(k-1)}\right\},$$

*where $R_0$ denotes the distance from initial point $w_0$ to solution set $X_*$, i.e.,*

$$R_0 = \|w_0 - x_*\| = \min\{\|w_0 - x\| : x \in X_*\}.$$

The following lemma provides a bound on the complexity of Algorithm 1 to satisfy (29), which connects the "inner" and "outer" perspectives.

**Proposition 5.3.** *Assume that $\lambda \geq 1/(L_f - \mu_f)$. Then in each call to ACG in Step **2** of Algorithm 1, after at most*

$$1 + \left\lceil \min\left\{2\sqrt{10\sigma^{-1}\lambda(L_f - \mu_f)}, \left(\frac{1}{4} + \frac{1}{2}\sqrt{\frac{2\lambda(L_f - \mu_f)}{1 + \lambda\mu_f}}\right)\ln\left(10\sigma^{-1}\lambda(L_f - \mu_f)\right)\right\}\right\rceil. \tag{71}$$

*ACG iterations, the condition* (29) *is satisfied.*

We are now ready to prove Theorem 2.3.

**Proof of Theorem 2.3:** Recall that by Proposition 5.3, the inner complexity of Algorithm 1 in Step **2** is

$$\tilde{\mathcal{O}}(1 + \sqrt{\lambda(L_f - \mu_f)}), \tag{72}$$

and by Proposition 5.2, the outer complexity of Algorithm 2 to find an $\varepsilon$-solution is

$$\tilde{\mathcal{O}}\left(1 + \min\left\{\frac{R_0}{\sqrt{\lambda\varepsilon}}, \frac{1}{\sqrt{\mu_f\lambda}}\right\}\right). \tag{73}$$

a) In the case $\mu_f = 0$, the outer complexity is $\tilde{\mathcal{O}}(1 + R_0/\sqrt{\lambda\varepsilon})$, hence the total complexity is

$$\tilde{\mathcal{O}}\left(\left(1 + \sqrt{\lambda L_f}\right)\left(1 + \frac{R_0}{\sqrt{\lambda\varepsilon}}\right)\right),$$

which becomes $\tilde{\mathcal{O}}\left(R_0\sqrt{L_f/\varepsilon}\right)$ under the assumption that $1/L_f \leq \lambda \leq R_0^2/\varepsilon$.

b) In the case $\mu_f > 0$, the total complexity immediately follows from (72), (73), and the assumption that $1/(L_f - \mu_f) \leq \lambda \leq \min\{1/\mu_f, R_0^2/\varepsilon\}$. ∎

## 5.2 Proof of Theorem 3.1

We consider two perspectives to prove Theorem 3.1: "inner" and "outer". First, we bound the number of inner iterations needed to satisfy the termination criterion $\|\mathcal{G}_{\mathcal{L}_\rho(\cdot,\lambda_k)}^{(2L+\mu)^{-1}}\| \leq \varepsilon_k/2$ in Step **1** of Algorithm 3. The inner bound is a direct result of the ACG convergence rate in Lemma 2.1, and we therefore defer the proof to Appendix E.

**Proposition 5.4.** *The number of ACG iterations required in the call to Algorithm 1 in Step **1** of Algorithm 3 is at most*

$$\tilde{\mathcal{O}}\left(1 + \frac{D(\sqrt{L_f} + \sqrt{\rho}\|A\|)}{\sqrt{\sigma\rho\varepsilon}}\right). \tag{74}$$

To bound the outer complexity, we will show that Algorithm 3 implements the LOrA framework, i.e., Algorithm 5, with the correspondence

$$\begin{aligned} \Phi(\cdot) = -d(\cdot), \quad \Gamma_k^{\mathrm{L}}(\cdot) = -\mathcal{L}(x_{k+1}, \cdot) + \frac{1}{2\rho}\|\cdot - \lambda_k\|^2, \quad \mu_{\mathrm{L}} = 0, \quad \lambda_{\mathrm{L}} = \rho, \quad \sigma_{\mathrm{L}} = \sigma; \\ \mathcal{A}_k^{\mathrm{L}} = \infty, \quad \delta_k^{\mathrm{L}} = \varepsilon_0 \alpha^k, \quad y_k^{\mathrm{L}} = x_k^{\mathrm{L}} = \lambda_k, \quad \hat{u}_{k+1}^{\mathrm{L}} = 0, \quad \alpha_{\mathrm{L}} = \alpha. \end{aligned} \tag{75}$$

We begin by showing that our choice of $\Gamma_k^{\mathrm{L}}$, $x_k^{\mathrm{L}}$, and $\hat{u}_k^{\mathrm{L}}$ satisfy (50) and (52).

**Lemma 5.5.** *Consider the sequences $\{\lambda_{k+1}\}$ and $\{x_{k+1}\}$ produced by Algorithm 3. Then, for every $k \geq 0$, the following statements hold:*

a) *for every $\nu \in \mathbb{R}^m$, we have*

$$-\mathcal{L}(x_{k+1}, \nu) + \frac{1}{2\rho}\|\nu - \lambda_k\|^2 \leq -d(\nu) + \frac{1}{2\rho}\|\nu - \lambda_k\|^2; \tag{76}$$

b)

$$\lambda_{k+1} = \operatorname*{argmin}_{\nu \in \mathbb{R}^m} \left\{ -\mathcal{L}(x_{k+1}, \nu) + \frac{1}{2\rho}\|\nu - \lambda_k\|^2 \right\}. \tag{77}$$

*Moreover, in light of* (75), (76) *and* (77) *correspond to* (50) *and* (52), *respectively.*

We now prove that on all iterations of Algorithm 3, either the inequality (51) holds or $(x_{k+1}, \lambda_{k+1})$ is an $\varepsilon$-primal-dual solution to (2) and the outer loop terminates. Combined with Lemma 5.5, we therefore guarantee that, until termination, Algorithm 3 is an instance of the LOrA framework (i.e., Algorithm 5).

**Proposition 5.6.** *For every $k \geq 0$, we have either*

$$-d(\lambda_{k+1}) + \frac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 - \Gamma_k^{\mathrm{L}}(\lambda_{k+1}) \leq \varepsilon_0 \alpha^k + \frac{\sigma}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2, \tag{78}$$

*which corresponds to* (51) *in view of* (75), *or $(x_{k+1}, \lambda_{k+1})$ is an $\varepsilon$-primal-dual solution to* (2).

**Proof**: Setting $L$ and $\mu$ as in (33) and applying Proposition B.2 to (2) with $(\tilde{x}, x^+, \lambda, \lambda^+) = (\tilde{x}_k, x_{k+1}, \lambda_k, \lambda_{k+1})$ and $\eta = (2L + \mu)^{-1}$, then the inner termination condition in Step **1**, i.e., $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\mathcal{L}_\rho(\cdot, \lambda_k)}(\tilde{x}_k)\| \leq \varepsilon_k/(2D)$, implies that there exists some $v \in \partial \mathcal{L}(\cdot, \lambda_{k+1})(x_{k+1})$ satisfying $\|v\| \leq \varepsilon_k/D$. It then follows by the definition of $\Gamma_k$ in (75), the Cauchy-Schwarz inequality, and Assumption 2(d) that we have

$$\begin{aligned} -d(\lambda_{k+1}) + \frac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 - \Gamma_k^{\mathrm{L}}(\lambda_{k+1}) &\overset{(75)}{=} \mathcal{L}(x_{k+1}, \lambda_{k+1}) - d(\lambda_{k+1}) \\ &\leq \langle v, x_{k+1} - u(\lambda_{k+1})\rangle \leq \|v\|\|x_{k+1} - u(\lambda_{k+1})\| \leq \varepsilon_k = \frac{\varepsilon_0 \alpha^k}{2} + \frac{\sigma\rho\varepsilon^2}{2}, \end{aligned} \tag{79}$$

where the first inequality follows from $v \in \partial \mathcal{L}(\cdot, \lambda_{k+1})(x_{k+1})$ and $u(\lambda_{k+1}) = \operatorname*{argmin}_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda_{k+1})$, and the last identity follows from the choice of $\varepsilon_k$ in Step **1**.

We now consider three cases to prove the proposition: 1) if $\|Ax_{k+1} - b\| \geq \varepsilon$, then we show (78) holds; 2) if $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\mathcal{L}_\rho(\cdot, \lambda_k)}(\tilde{x}_k)\| \geq \varepsilon/2$, then we show (78) holds; and 3) if both conditions are violated, then we show that $(x_{k+1}, \lambda_{k+1})$ is an $\varepsilon$-primal-dual solution to (2).

**Case 1)** If $\|Ax_{k+1}-b\| \geq \varepsilon$, then $\rho\varepsilon^2 \leq \rho\|Ax_{k+1}-b\|^2 \overset{(34)}{=} \rho^{-1}\|\lambda_{k+1}-\lambda_k\|^2$. Then (79) and (34) imply that

$$-d(\lambda_{k+1}) + \frac{1}{2\rho}\|\lambda_{k+1}-\lambda_k\|^2 - \Gamma_k^{\mathrm{L}}(\lambda_{k+1}) \overset{(79)}{\leq} \frac{\varepsilon_0\alpha^k}{2} + \frac{\sigma\rho\varepsilon^2}{2} \leq \frac{\varepsilon_0\alpha^k}{2} + \frac{\sigma}{2\rho}\|\lambda_k-\lambda_{k+1}\|^2,$$

which satisfies (78).

**Case 2)** If $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\mathcal{L}_\rho(\cdot,\lambda_k)}(\tilde{x}_k)\| \geq \varepsilon/2$, then the termination condition of the inner solver, i.e., $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\mathcal{L}_\rho(\cdot,\lambda_k)}(\tilde{x}_k)\| \leq \varepsilon_k/(2D)$, implies $\varepsilon_k \geq D\varepsilon$. By the condition on $\sigma$ in Algorithm 3, we have $\sigma\rho\varepsilon \leq D/2$. Then,

$$D\varepsilon \leq \varepsilon_k = \frac{\varepsilon_0\alpha^k}{2} + \frac{\sigma\rho\varepsilon^2}{2} \leq \frac{\varepsilon_0\alpha^k}{2} + \frac{\varepsilon D}{4},$$

which implies

$$\frac{\varepsilon_0\alpha^k}{2} \geq \frac{3D\varepsilon}{4} \geq \frac{\sigma\rho\varepsilon^2}{2}. \tag{80}$$

Thus, by (79), we obtain

$$-d(\lambda_{k+1}) + \frac{1}{2\rho}\|\lambda_{k+1}-\lambda_k\|^2 - \Gamma_k^{\mathrm{L}}(\lambda_{k+1}) \overset{(79)}{\leq} \frac{\varepsilon_0\alpha^k}{2} + \frac{\sigma\rho\varepsilon^2}{2} \overset{(80)}{\leq} \varepsilon_0\alpha^k,$$

which satisfies (78).

**Case 3)** We now consider the third case, where the conditions for the first two cases fail to hold, that is, Algorithm 3 terminates in Step **3**. Now that $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\mathcal{L}_\rho(\cdot,\lambda_k)}(\tilde{x}_k)\| \leq \varepsilon/2$, then Proposition B.2 with $(\tilde{x}, x^+, \lambda, \lambda^+) = (\tilde{x}_k, x_{k+1}, \lambda_k, \lambda_{k+1})$ and $\eta = (2L+\mu)^{-1}$ implies that there exists a $v \in \partial\mathcal{L}(\cdot,\lambda_{k+1})(x_{k+1})$ satisfying $\|v\| \leq \varepsilon$. Hence, $\|Ax_{k+1}-b\| \leq \varepsilon$ and $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\mathcal{L}_\rho(\cdot,\lambda_k)}(\tilde{x}_k)\| \leq \varepsilon/2$ indicates that $(x_{k+1},\lambda_{k+1})$ is an $\varepsilon$-primal-dual solution by (4).

Therefore, we complete the proof. ∎

We are now ready to prove Theorem 3.1.

**Proof of Theorem 3.1:** Recall from Proposition 5.4 that the inner complexity to satisfy the inner termination condition $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\mathcal{L}_\rho(\cdot,\lambda_k)}(\tilde{x}_k)\| \leq \varepsilon_k/(2D)$ is as in (74).

Observe that $\sigma = 1/2$ and $\rho = \varepsilon^{-1}$ satisfy the requirement $2\sigma\rho = 1/\varepsilon \leq D/\varepsilon$ (see initialization in Algorithm 3) in view of Assumption 2(d). Set $L$ and $\mu$ as in (33). By the inner termination condition, our choice $\varepsilon_0 = \varepsilon$, and the condition $\rho\sigma \leq D/(2\varepsilon)$, for all iterations $k$ we have

$$\|\mathcal{G}^{(2L+\mu)^{-1}}_{\mathcal{L}_\rho(\cdot,\lambda_k)}(\tilde{x}_k)\| \leq \frac{\varepsilon_k}{2D} \leq \frac{\varepsilon_0}{4D} + \frac{\rho\sigma\varepsilon^2}{4D} \leq \frac{\varepsilon}{2}.$$

Then, Proposition B.2 applied to problem (2) with $(\tilde{x}, x^+, \lambda, \lambda^+) = (\tilde{x}_k, x_{k+1}, \lambda_k, \lambda_{k+1})$ and $\eta = (2L+\mu)^{-1}$ implies that for all iterations $k \geq 0$, there exists a $v \in \partial\mathcal{L}(\cdot,\lambda_{k+1})(x_{k+1})$ satisfying $\|v\| \leq \varepsilon$.

By Lemma 5.5 and Proposition 5.6, Algorithm 3 is an instance of the LOrA framework (i.e., Algorithm 5) until termination with the correspondence (75). Then using Theorem 4.4 with $R_0^{\mathrm{L}} = R_\Lambda$, we have

$$\rho \min_{1\leq i\leq k}\|Ax_i - b\| \overset{(34)}{=} \min_{1\leq i\leq k}\|\lambda_i - \lambda_{i-1}\| \overset{(75)}{=} \min_{1\leq i\leq k}\|y_i^{\mathrm{L}} - x_{i-1}^{\mathrm{L}}\| \overset{(59),(75)}{\leq} \frac{R_\Lambda}{\sqrt{1-\sigma}\sqrt{k}} + \sqrt{\frac{2\rho\bar{\delta}_k^{\mathrm{L}}}{1-\sigma}}.$$

It follows from the definition of $\bar{\delta}_k^{\mathrm{L}}$ in Theorem 4.4 and $\delta_k^{\mathrm{L}} = \varepsilon_0 \alpha^k = \varepsilon \alpha^k$ from (75) that

$$\bar{\delta}_k^{\mathrm{L}} = \frac{\sum_{i=0}^{k-1} \varepsilon \alpha^i}{k} \leq \frac{\varepsilon}{(1-\alpha)k}.$$

The above two inequalities immediately imply the outer complexity to guarantee near feasibility $\min_{1 \leq i \leq k} \|Ax_i - b\| \leq \varepsilon$ is

$$\mathcal{O}\left(1 + \frac{R_\Lambda^2 + \rho\varepsilon}{(1-\sigma)\rho^2\varepsilon^2}\right) \tag{81}$$

outer iterations, since at least one outer iteration is needed to ensure stationarity.

Combining the inner complexity from (74) and outer complexity from (81), and substituting $\rho = \varepsilon^{-1}$, we obtain the total complexity as in (35). ∎

### 5.3 Proof of Theorem 3.4

We prove Theorem 3.4 by following the same approach as in Subsection 5.2. First, we will provide a bound on the inner complexity in each call to Algorithm 1 in Step **2**, then we will bound the outer complexity by proving that Algorithm 4 is an instance of FLOrA (i.e., Algorithm 6). However, the inclusion of dual perturbations requires more care in the outer analysis than in the prior subsection. Our choice of the auxiliary point $z_{k+1}^{\mathrm{F}}$ in the FLOrA analysis will play a crucial role in our argument.

Before providing complexity bounds, we show that our primal-dual perturbations in (38) do not add dependence on $\varepsilon^{-1}$, as observed for primal-only perturbations in [26, Appendix A]. Instead, when both $\gamma_p$ and $\gamma_d$ are $\mathcal{O}(\varepsilon)$, the dependence on $\varepsilon^{-1}$ disappears. The proof of the following lemma is deferred to Appendix E.2.2.

**Lemma 5.7.** *Let $\Lambda_* = \{\lambda : d(\lambda) = d_*\}$ be the set of optimal multipliers for the original problem (2). Define $R_\Lambda := \|\lambda_0 - \lambda_*\| = \min\{\|\lambda_0 - \lambda\| : \lambda \in \Lambda_*\}$ and $R_{\tilde{\Lambda}} := \|\tilde{\lambda}_* - \lambda_0\|$ where $\tilde{\lambda}_*$ is the unique minimizer of $-\tilde{d}(\cdot)$ and $-\tilde{d}(\cdot)$ is as in (38). Suppose $\gamma_p = \varepsilon/(2D)$ and $\gamma_d = C_0\varepsilon/(R_{\tilde{\Lambda}})$ for some $C_0 > 0$, then we have*

$$R_{\tilde{\Lambda}} \leq R_\Lambda + \frac{D}{4C_0}. \tag{82}$$

With the perturbed bound proven, we proceed with our proof of Theorem 3.4 by bounding the inner complexity in Step **2**. The proof is nearly identical to that of Proposition 5.4, and is likewise deferred to Appendix E.2.3.

**Proposition 5.8.** *Choosing $\gamma_p = \varepsilon/(2D)$, then the number of ACG iterations required in the call to Algorithm 1 in Step 2 of Algorithm 4 is at most*

$$\tilde{\mathcal{O}}\left(1 + \frac{\sqrt{D}(\sqrt{L_f} + \sqrt{\rho}\|A\|)}{\sqrt{\varepsilon}}\right). \tag{83}$$

We now switch to the "outer" perspective. Define the point

$$\hat{\lambda}_{k+1} = \frac{\lambda_{k+1}}{1 + \gamma_d \rho} \tag{84}$$

and the function

$$\Gamma_k^\lambda(\cdot) = -\tilde{\mathcal{L}}(x_{k+1}, \lambda_{k+1}) + \frac{1}{2\rho}\|\lambda_{k+1} - \tilde{\nu}_k\|^2 + \langle \gamma_d \lambda_{k+1}, \cdot - \lambda_{k+1}\rangle + \frac{1 + \gamma_d\rho}{2\rho}\|\cdot - \lambda_{k+1}\|^2, \tag{85}$$

which is a $(\rho^{-1}+\gamma_d)$-strongly convex approximation of $-\tilde{\mathcal{L}}(x_{k+1},\cdot)+\|\cdot-\tilde{\nu}_k\|^2/(2\rho)$ at $\lambda_{k+1}$.

We will show that Algorithm 4 is an instance of the FLOrA framework with the correspondence

$$\begin{gathered}\Phi(\cdot)=-\tilde{d}(\cdot),\ \Gamma_k^{\mathrm{F}}(\cdot)=\Gamma_k^{\lambda}(\cdot),\ \mathcal{A}_k^{\mathrm{F}}=\infty,\ \alpha_{\mathrm{F}}=\alpha,\ \mu_{\mathrm{F}}=\gamma_d,\ \sigma_{\mathrm{F}}=\sigma,\ \lambda_{\mathrm{F}}=\rho;\\ \delta_k^{\mathrm{F}}=\varepsilon_0\alpha^k,\ y_k^{\mathrm{F}}=\tilde{y}_k^{\mathrm{F}}=\lambda_k,\ z_k^{\mathrm{F}}=\hat{\lambda}_k,\ x_k^{\mathrm{F}}=\nu_k,\ \tilde{x}_k^{\mathrm{F}}=\tilde{\nu}_k,\ u_k^{\mathrm{F}}=\rho^{-1}(\tilde{\nu}_{k-1}-\lambda_k),\ \hat{u}_k^{\mathrm{F}}=0.\end{gathered}\tag{86}$$

First, we show that the conditions (63), (65), and (66) are satisfied, along with a summability bound related to the absolute error sequence $\{\varepsilon_0\alpha^k\}$.

**Lemma 5.9.** *The following statements hold for every $k\geq 0$,*

a) *for every $\nu\in\mathbb{R}^m$*

$$\Gamma_k^{\lambda}(\nu)\leq -\tilde{d}(\nu)+\frac{1}{2\rho}\|\nu-\tilde{\nu}_k\|^2;$$

b) $\hat{\lambda}_{k+1}=\operatorname*{argmin}_{\nu\in\mathbb{R}^m}\Gamma_k^{\lambda}(\nu)$ *and*

$$\min_{\nu\in\mathbb{R}^m}\Gamma_k^{\lambda}(\nu)=-\tilde{\mathcal{L}}(x_{k+1},\lambda_{k+1})+\frac{1}{2\rho}\|\lambda_{k+1}-\tilde{\nu}_k\|^2-\frac{\gamma_d^2\rho}{2(1+\gamma_d\rho)}\|\lambda_{k+1}\|^2.\tag{87}$$

c) *letting $u_{k+1}=\rho^{-1}(\tilde{\nu}_k-\hat{\lambda}_{k+1})$, we can rewrite* (43) *as*

$$\nu_{k+1}=\frac{1}{\tau_{k+1}}\left(\tau_k\nu_k+b_k\gamma_d\hat{\lambda}_{k+1}-b_ku_{k+1}\right).$$

d) *defining $\beta=\sqrt{\alpha}(1+\sqrt{\rho\gamma_d})<1$, we have $C\leq\rho(1-\beta)^{-4}<\infty$, where $C$ is as in* (44).

*Moreover, in light of* (86)*, statements a), b), and c) correspond to* (63), (65), *and* (66)*, respectively, and therefore* (40) *is equivalent to* (62).

Analyzing Algorithm 4 as an instance of Algorithm 6 now requires that we show (64) holds with the correspondence (86). The following proposition is the analogue of Proposition 5.6 from the prior subsection, retaining the same "three-case" structure while adapting the analysis to the dual perturbations.

**Proposition 5.10.** *Suppose $\gamma_d>0$ satisfies*

$$\gamma_d\leq\min\left\{\frac{\sqrt{\sigma}}{2\sqrt{3}\rho},\frac{\sqrt{\sigma}\varepsilon}{4\sqrt{3}\mathcal{R}}\right\},\tag{88}$$

*where $\mathcal{R}$ is as in* (44)*. Then, for every $k\geq 0$, we have either*

$$-\tilde{d}(\lambda_{k+1})+\frac{1}{2\rho}\|\lambda_{k+1}-\tilde{\nu}_k\|^2-\Gamma_k^{\lambda}(\hat{\lambda}_{k+1})\leq\frac{\sigma}{2\rho}\|\lambda_{k+1}-\tilde{\nu}_k\|^2+\varepsilon_0\alpha^k,\tag{89}$$

*which corresponds to* (64) *in view of* (86)*, or $(x_{k+1},\lambda_{k+1})$ is an $\varepsilon$-primal-dual solution to* (2).

**Proof**: We prove the proposition by induction. Throughout, let $L$ and $\mu$ be as in (41). First, we note that Lemma 5.9(b) and the definition of $\Gamma_k^\lambda$ in (85) imply that

$$-\tilde{d}(\lambda_{k+1}) + \frac{1}{2\rho}\|\lambda_{k+1} - \tilde{\nu}_k\|^2 - \Gamma_k^\lambda(\hat{\lambda}_{k+1}) \overset{(87)}{=} \tilde{\mathcal{L}}(x_{k+1}, \lambda_{k+1}) - \tilde{d}(\lambda_{k+1}) + \frac{\gamma_d^2\rho}{2(\gamma_d\rho+1)}\|\lambda_{k+1}\|^2. \quad (90)$$

Applying Proposition B.2 to (37) with $(\tilde{x}, x^+, \lambda, \lambda^+) = (\tilde{x}_k, x_{k+1}, \tilde{\nu}_k, \lambda_{k+1})$, $\eta = (2L+\mu)^{-1}$, and $f$ replaced by $f + \gamma_p\|\cdot - x_0\|^2/2$, then the inner termination condition in Step **2**, i.e., $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\tilde{\mathcal{L}}_\rho(\cdot,\tilde{\nu}_k)}(\tilde{x}_k)\| \le \varepsilon_k/2D$, implies that there exists a subgradient $v \in \partial\tilde{\mathcal{L}}(\cdot, \lambda_{k+1})(x_{k+1})$ satisfying $\|v\| \le \varepsilon_k/D$. It then follows from the Cauchy-Schwarz inequality and Assumption 2(d) that

$$\tilde{\mathcal{L}}(x_{k+1}, \lambda_{k+1}) - \tilde{d}(\lambda_{k+1}) \le \langle v, x_{k+1} - \tilde{u}(\lambda_{k+1})\rangle \le \|v\|D \le \varepsilon_k = \frac{7\varepsilon_0\alpha^k}{8} + \frac{\sigma\rho\varepsilon^2}{8}, \quad (91)$$

where the first inequality follows from $v \in \partial\tilde{\mathcal{L}}(\cdot, \lambda_{k+1})(x_{k+1})$ and $\tilde{u}(\lambda_{k+1}) = \operatorname{argmin}_{x\in\mathbb{R}^n} \tilde{\mathcal{L}}(x, \lambda_{k+1})$, and the last identity follows from the choice of $\varepsilon_k$ in Step **1**.

By our choice $\lambda_0 = \nu_0 = 0$, for the base case we have $\tilde{\nu}_0 = 0$ and so $\|\lambda_1\| = \|\tilde{\nu}_0 - \lambda_1\|$. It follows from the initialization in Algorithm 4 that $\sigma\rho\varepsilon^2 \le \varepsilon/4 \le \varepsilon_0/4$, which together with (91) implies that

$$\tilde{\mathcal{L}}(x_1, \lambda_1) - \tilde{d}(\lambda_1) + \frac{\gamma_d^2\rho}{2(\gamma_d\rho+1)}\|\lambda_1\|^2 \overset{(91)}{\le} \frac{7\varepsilon_0}{8} + \frac{\sigma\rho\varepsilon^2}{8} + \frac{\sigma}{2\rho}\|\tilde{\nu}_0 - \lambda_1\|^2 \le \varepsilon_0 + \frac{\sigma}{2\rho}\|\tilde{\nu}_0 - \lambda_1\|^2.$$

In view of (90), the above inequality proves (89) with $k = 0$, which is the base case of the proposition.

Now we assume the proposition holds for iterations $0 \le n \le k-1$. Without loss of generality, we assume that (89) holds with $k$ replaced by $k-1$, otherwise $(x_k, \lambda_k)$ is already an $\varepsilon$-primal-dual solution to (2). Hence, (89) and Lemma 5.9 imply that Algorithm 4 is an instance of the FLOrA framework (i.e., Algorithm 6) under the correspondence (86). Then, using Lemma C.6 with $\mathcal{R}_{\mathrm{F}} = \mathcal{R}$ (which is defined in (44)) and the correspondence (86), we have

$$\|\tilde{\nu}_k - \tilde{\lambda}_*\| \overset{(86)}{=} \|\tilde{x}_k^{\mathrm{F}} - x_*\| \overset{(121)}{\le} \mathcal{R}. \quad (92)$$

It thus follows from the triangle inequality and the Cauchy-Schwarz inequality that

$$\begin{aligned}\|\lambda_{k+1}\|^2 &\le (\|\lambda_{k+1} - \tilde{\nu}_k\| + \|\tilde{\nu}_k - \tilde{\lambda}_*\| + \|\tilde{\lambda}_*\|)^2 \\ &\le 3(\|\lambda_{k+1} - \tilde{\nu}_k\|^2 + \|\tilde{\nu}_k - \tilde{\lambda}_*\|^2 + \|\tilde{\lambda}_*\|^2) \overset{(92)}{\le} 3\left(\|\lambda_{k+1} - \tilde{\nu}_k\|^2 + 2\mathcal{R}^2\right),\end{aligned}$$

where the last inequality is due to (92) and the fact that $\|\tilde{\lambda}_*\| = \|\tilde{\lambda}_* - \lambda_0\| \le \hat{R}_{\tilde{\Lambda}} \le \mathcal{R}$ in view of (44). The above inequality and the requirement on $\gamma_d$ in (88) further imply that

$$\frac{\gamma_d^2\rho}{2(\gamma_d\rho+1)}\|\lambda_{k+1}\|^2 \le \frac{3\gamma_d^2\rho}{2}\left(\|\lambda_{k+1} - \tilde{\nu}_k\|^2 + 2\mathcal{R}^2\right) \overset{(88)}{\le} \frac{\sigma}{8\rho}\|\lambda_{k+1} - \tilde{\nu}_k\|^2 + \frac{\sigma\rho\varepsilon^2}{16}. \quad (93)$$

Putting together (90), (91), and (93), we obtain

$$-\tilde{d}(\lambda_{k+1}) + \frac{1}{2\rho}\|\lambda_{k+1} - \tilde{\nu}_k\|^2 - \Gamma_k^\lambda(\hat{\lambda}_{k+1}) \le \frac{7\varepsilon_0\alpha^k}{8} + \frac{3\sigma\rho\varepsilon^2}{16} + \frac{\sigma}{8\rho}\|\lambda_{k+1} - \tilde{\nu}_k\|^2. \quad (94)$$

We now consider three cases to prove the proposition: 1) if $\|Ax_{k+1} - b\| \ge \varepsilon$, then we show that (89) holds; 2) if $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\tilde{\mathcal{L}}_\rho(\cdot,\tilde{\nu}_k)}(\tilde{x}_k)\| \ge \varepsilon/4$, then we show (89) holds; and 3) if both conditions are violated, then we show that $(x_{k+1}, \lambda_{k+1})$ is an $\varepsilon$-primal-dual solution to (2).

**Case 1)** Since $\|Ax_{k+1} - b\| \geq \varepsilon$, it follows from (42) that

$$\frac{3\sigma\rho\varepsilon^2}{16} \leq \frac{3\sigma\rho}{16}\|Ax_{k+1} - b\|^2 \overset{(42)}{=} \frac{3\sigma}{16\rho}\|\lambda_{k+1} - \tilde{\nu}_k\|^2,$$

which together with (94) implies that

$$-\tilde{d}(\lambda_{k+1}) + \frac{1}{2\rho}\|\lambda_{k+1} - \tilde{\nu}_k\|^2 - \Gamma_k^{\lambda}(\hat{\lambda}_{k+1}) \leq \frac{7\varepsilon_0\alpha^k}{8} + \frac{5\sigma}{16\rho}\|\lambda_{k+1} - \tilde{\nu}_k\|^2.$$

Hence, (89) immediately follows.

**Case 2)** Since $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\tilde{\mathcal{L}}_\rho(\cdot,\tilde{\nu}_k)}(\tilde{x}_k)\| \geq \varepsilon/4$, the inner termination condition in Step **2** of Algorithm 4 implies that $\varepsilon_k \geq D\varepsilon/2$. It thus follows from the choice of $\varepsilon_k$ in Step **1** that

$$\frac{D\varepsilon}{2} \leq \varepsilon_k = \frac{7\varepsilon_0\alpha^k}{8} + \frac{\sigma\rho\varepsilon^2}{8} \leq \frac{7\varepsilon_0\alpha^k}{8} + \frac{D\varepsilon}{32},$$

where the inequality is due to $\sigma\rho \leq 1/(4\varepsilon) \leq D/(4\varepsilon)$ by the initialization of Algorithm 4 and Assumption 2(d). The above inequality thus indicates that

$$\frac{7\varepsilon_0\alpha^k}{8} \geq \frac{15D\varepsilon}{32} \geq \frac{15\sigma\rho\varepsilon^2}{8} \implies \frac{3\sigma\rho\varepsilon^2}{16} \leq \frac{7\varepsilon_0\alpha^k}{80}.$$

Plugging the above bound into (94), we obtain

$$-\tilde{d}(\lambda_{k+1}) + \frac{1}{2\rho}\|\lambda_{k+1} - \tilde{\nu}_k\|^2 - \Gamma_k^{\lambda}(\hat{\lambda}_{k+1}) \leq \frac{77\varepsilon_0\alpha^k}{80} + \frac{\sigma}{8\rho}\|\lambda_{k+1} - \tilde{\nu}_k\|^2.$$

Hence, (89) immediately follows.

**Case 3)** We now consider the third case, where the conditions for the first two cases fail to hold, that is, Algorithm 4 terminates in Step **4**. Now that $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\tilde{\mathcal{L}}_\rho(\cdot,\tilde{\nu}_k)}(\tilde{x}_k)\| \leq \varepsilon/4$, then Proposition B.2 applied to problem (37) with $(\tilde{x}, x^+, \lambda, \lambda^+) = (\tilde{x}_k, x_{k+1}, \tilde{\nu}_k, \lambda_{k+1})$, $\eta = (2L+\mu)^{-1}$, and $f$ replaced by $f(\cdot) + \gamma_p\|\cdot - x_0\|^2/2$ implies that there exists a $v \in \partial\tilde{\mathcal{L}}(\cdot, \lambda_{k+1})(x_{k+1})$ such that $\|v\| \leq \varepsilon/2$. Using the initialization $\gamma_p = \varepsilon/(2D)$ in Algorithm 4, Lemma 3.3 with $(x, \lambda) = (x_{k+1}, \lambda_{k+1})$ implies that there exists a $v' \in \partial\mathcal{L}(\cdot, \lambda_{k+1})(x_{k+1})$ such that $\|v'\| \leq \varepsilon$. Hence $\|Ax_{k+1} - b\| \leq \varepsilon$ and $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\tilde{\mathcal{L}}_\rho(\cdot,\tilde{\nu}_k)}(\tilde{x}_k)\| \leq \varepsilon/4$ indicates that $(x_{k+1}, \lambda_{k+1})$ is an $\varepsilon$-primal-dual solution by (4).

Therefore, we finish the inductive proof and thus complete the proof of the lemma. ∎

We are now ready to prove Theorem 3.4.

**Proof of Theorem 3.4:** Recall from Proposition 5.8 that the inner complexity to satisfy the inner termination condition $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\tilde{\mathcal{L}}_\rho(\cdot,\tilde{\nu}_k)}(\tilde{x}_k)\| \leq \varepsilon_k/(2D)$ is as in (83). Then, we simply need to bound the complexity of the outer loop.

It is trivial to show that the parameter choices satisfy the conditions in the initialization of Algorithm 4,

$$4\rho\sigma\varepsilon \leq 1, \quad \varepsilon_0 \geq \varepsilon, \quad 0 \leq \alpha < (1 + \sqrt{\gamma_d\rho})^{-2}. \tag{95}$$

We then proceed to bound the outer iteration complexity to satisfy each of the termination criteria in Step **4** of Algorithm 4. Combining the outer complexity with the inner complexity in (83) will then yield the total complexity in (46).

First, we bound the complexity to satisfy the stationarity condition $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\tilde{\mathcal{L}}_\rho(\cdot,\tilde{\nu}_k)}(\tilde{x}_k)\| \le \varepsilon/4$, where $L$ and $\mu$ are as in (41). By the inner termination condition in Step **2**, i.e., $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\tilde{\mathcal{L}}_\rho(\cdot,\tilde{\nu}_k)}(\tilde{x}_k)\| \le \varepsilon_k/(2D)$, the inequality $\|\mathcal{G}^{(2L+\mu)^{-1}}_{\tilde{\mathcal{L}}_\rho(\cdot,\tilde{\nu}_k)}(\tilde{x}_k)\| \le \varepsilon/4$ is satisfied by any iteration with $\varepsilon_k \le D\varepsilon/2$. By Assumption 2(d) and (95), we obtain $\sigma\rho\varepsilon^2 \le \varepsilon/4 \le D\varepsilon/4$. Then, from the choice of $\varepsilon_k$ in Step **1** of Algorithm 4, the condition

$$\varepsilon_k = \frac{7}{8}\varepsilon_0\alpha^k + \frac{\sigma\rho\varepsilon^2}{8} \le \frac{7}{8}\varepsilon_0\alpha^k + \frac{D\varepsilon}{32} \le \frac{D\varepsilon}{2}$$

is satisfied when $7\varepsilon_0\alpha^k/8 \le 15D\varepsilon/32$, which occurs in

$$N_\alpha = \left\lceil \frac{\log(15D\varepsilon/32) - \log(7\varepsilon_0/8)}{\log\alpha} \right\rceil \le 1 + \frac{\log(15D\varepsilon/32) - \log(7\varepsilon_0/8)}{\log\alpha} \overset{(45)}{\le} 1 + \sqrt{\frac{D}{\rho\varepsilon}} \tag{96}$$

outer iterations, where the second inequality follows by the second condition on $\alpha$ in (45). Therefore, $N_\alpha = \mathcal{O}(1 + \sqrt{D/(\rho\varepsilon)})$.

Next, we bound the outer iteration complexity required to satisfy the termination condition $\|Ax_k - b\| \le \varepsilon$. Combining $\sigma\varepsilon \le 1/(4\rho)$ from (95) with the choice of $\gamma_d$ and $\sigma = 1/4$, we can show that the condition (88) in Proposition 5.10 is satisfied. Therefore, Lemma 5.9 and Proposition 5.10 imply that Algorithm 4 is an instance of the FLOrA framework (i.e., Algorithm 6) with the correspondence (86).

Then, using Theorem 4.5 with $R_0^{\mathrm{F}} = \|\tilde{\lambda}_*\| \le \hat{R}_{\tilde{\Lambda}}$ (see (44)) and $C_{\mathrm{F}} = C$, we have

$$\rho\|Ax_k - b\| \overset{(42)}{=} \|\lambda_k - \tilde{\nu}_{k-1}\| \overset{(86)}{=} \|\tilde{y}_k^{\mathrm{F}} - \tilde{x}_{k-1}^{\mathrm{F}}\| \overset{(68),(86)}{\le} \frac{\sqrt{\rho}\hat{R}_{\tilde{\Lambda}} + \sqrt{2\rho\varepsilon_0 C}}{\sqrt{(1-\sigma)B_k}}.$$

Combining the above inequality with Lemma C.2(c) immediately implies the outer complexity to guarantee near feasibility $\|Ax_{k+1} - b\| \le \varepsilon$ is $\tilde{\mathcal{O}}(1 + 1/\sqrt{\rho\gamma_d})$. Using Lemma 5.9(d) with $\beta = \sqrt{9/10}$ from (45) and the choice $\varepsilon_0 = \rho^{-1}$, we can show that $\gamma_d = \mathcal{O}(\varepsilon/\hat{R}_{\tilde{\Lambda}})$ and $\mathcal{R} = \mathcal{O}(\hat{R}_{\tilde{\Lambda}})$. Then by Lemma 5.7, we have that $\mathcal{R} = \mathcal{O}(\hat{R}_{\tilde{\Lambda}}) = \mathcal{O}(\hat{R}_\Lambda + D)$.

Accordingly, the outer iteration count $k$ to satisfy $\|Ax_k - b\| \le \varepsilon$ is

$$\tilde{\mathcal{O}}\left(1 + \frac{1}{\sqrt{\rho\gamma_d}}\right) = \tilde{\mathcal{O}}\left(1 + \frac{\sigma^{3/4}\sqrt{\mathcal{R}}}{\sqrt{\rho\varepsilon}}\right) = \tilde{\mathcal{O}}\left(1 + \frac{\sqrt{\hat{R}_\Lambda + D}}{\sqrt{\rho\varepsilon}}\right), \tag{97}$$

which is of the same order as $N_\alpha$ in (96).

Combining the inner complexity from (83) and the outer complexity from (97), substituting $\rho = L_f/\|A\|^2$, and using $D \ge 1$, we obtain the total complexity as in (46). ∎

# 6 Numerical Experiments

In this section we provide numerical illustrations of the proposed primal and dual methods. All code is implemented in Julia and is publicly available[2]. Details of numerical experiments (problem generation, libraries, etc.) can be found in Appendix A. In-depth experimental analysis is beyond the scope of this work, and these tests should be taken as preliminary illustrations.

[2] https://github.com/mxburns2022/PrimalDualRestart

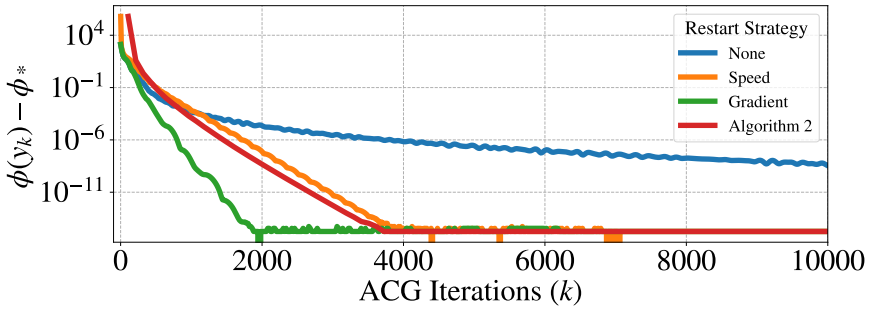



Figure 1: Numerical results for Restarted ACG algorithms.

### 6.1 Primal Methods: Restarted ACG

We compare Restarted ACG (Algorithm 2) to baseline ACG (Algorithm 1, "None") as well as two prominent restart schemes from literature: "gradient" restarting [37] and "speed" restarting [48].

Gradient restarting is a heuristic scheme that restarts the ACG solver whenever the gradient mapping forms an acute angle with the update direction, i.e., $\langle \tilde{x}_k - y_{k+1}, y_{k+1} - y_k \rangle > 0$.

For speed restarting, we restart the acceleration whenever the distance between adjacent iterates decreases, $\|y_{k+1} - y_k\| < \|y_k - y_{k-1}\|$, motivated by the continuous-time limit of ACG [48]. To prevent the speed scheme from restarting too often, we only allow restarts at most every $k_{\min}$ iterations. As in [48], we set $k_{\min} = 10$.

We focus on the sparse linear regression/LASSO problem [50]

$$\phi_* := \min_{x \in \mathbb{R}^n} \left\{ \phi(x) := \frac{1}{2}\|Ax - b\|_2^2 + \gamma \|x\|_1 \right\}. \tag{98}$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $\gamma > 0$. We use $n = 1000$, $m = 500$, and $\gamma = 1/2$.

Fig. 1 shows the estimated function value gap $\phi(y_k) - \phi_*$, where $\phi_*$ is the best solution found by any solver, versus the number of ACG iterations. All of the restart methods are significantly faster than baseline ACG ("None"). Speed restarting and Restarted ACG behave quite similarly, while gradient restarting has the most rapid convergence.

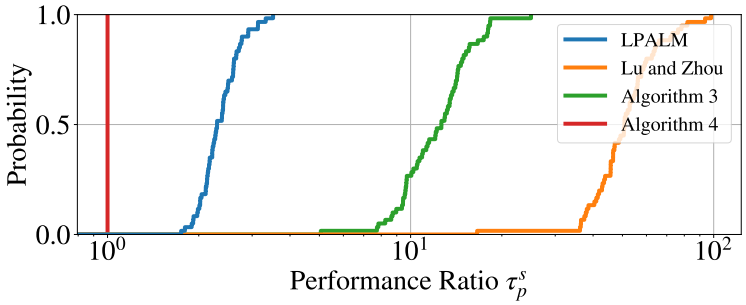
### 6.2 Dual Methods: Augmented Lagrangian

In this subsection, we compare several proposed ALM variants in linearly-constrained quadratic programming (LCQP). The LCQP problem is given by

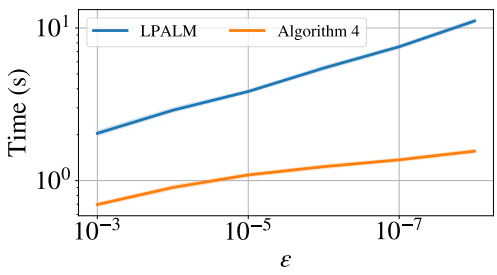
$$\hat{\phi}_* := \min_{x \in \mathbb{R}^n} \left\{ \phi(x) := \frac{1}{2} x^\top M x + c^\top x + \delta_{\mathrm{Q}}(x) : Ax = b \right\},$$

where $M \in \mathbb{R}^{n \times n}$ is positive semi-definite, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ is full rank, $b \in \mathbb{R}^m$, and $m \leq n$. For the nonsmooth component, we choose the indicator function $\delta_Q(\cdot)$, where $Q$ is the $n$-dimensional box with side length 20 centered at 0, $Q = \{x : -10 \leq x_i \leq 10 \text{ for all } 1 \leq i \leq n\}$. The LCQP problem is a staple of model predictive control [12] and as a subproblem in algorithms for nonsmooth optimization [6, Lemma 10.8], motivating its inclusion.

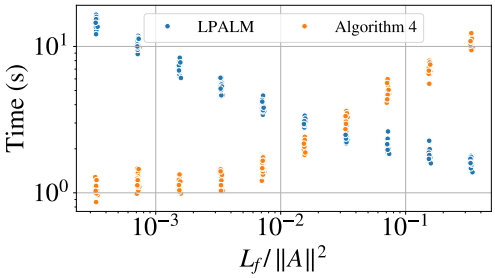
In addition to Algorithms 3 and 4, we compare against two $\tilde{\mathcal{O}}(\varepsilon^{-1})$ ALM proposals from the literature: the decreasing error/decreasing penalty scheme from [26, Algorithm 2] ("Lu and Zhou") and the linearized proximal ALM ("LPALM") with static $\rho$. LPALM is derived from the LPADMM method of [21, Algorithm 1] by setting one block to zero. The inner loop of Lu and Zhou is




(a) Performance profile of I-ALM algorithms in $n = 200$, $m = 100$ LCQP instances with $\varepsilon = 10^{-3}$




(b) Wall time scaling of Algorithm 4 and [21, Algorithm 1] with varying accuracy $\varepsilon^{-1}$.




(c) Wall time scaling of Algorithm 4 and [21, Algorithm 1] with varying $L_f/\|A\|^2$ with relative accuracy $L_f/\varepsilon$ and optimizer held constant.

Figure 2: Numerical experiments for ALM variants tested.

terminated based on the absolute error condition $\mathcal{L}_\rho(x_{k+1}, \lambda_k) - \min_{x\in\mathbb{R}^n} \mathcal{L}_\rho(x, \lambda_k) \leq \varepsilon_k$, which is estimated by the gradient mapping norm using Lemma B.1(a). We tested Algorithm 3 with a variety of $\rho$ values, finding that $\rho = 1.0$ was the most performant in practice. For Algorithm 4, we set $\rho = \sqrt{m}\|M\|/\|A\|^2$ (i.e., $\sqrt{m}L_f/\|A\|^2$).

Figure 2(a) shows the performance profile [10] of the four algorithms across 60 randomly generated LCQP instances with $n = 200$ and $m = 100$. We solve each instance to $\varepsilon = 10^{-3}$ accuracy (using the definition in (4)), terminating when an $\varepsilon$-primal-dual solution is detected. The x-axis is the "performance ratio" $\tau_p^s = t_p^s / \min_s t_p^s$, where $t_p^s$ is the elapsed wall-time needed for solver $s$ to reach the target accuracy on problem $p$, i.e., $\tau_p^s = 1$ if solver $s$ was first to achieve $\varepsilon \leq 10^{-3}$ on problem $p$. The y-axis of Figure 2(a) shows the cumulative distribution of $\tau_p^s$ for each solver across the 60 instances tested. Algorithm 4 shows a clear advantage, with LPALM placing second.

Focusing on the two most performant methods, Figs. 2(b) and 2(c) compare LPALM and Algorithm 4 across 20 LCQP instances with $n = 1000$ and $m = 500$. Fig 2(b) fixes the problem set ($L_f = 1.0$, $\|A\| \approx 17$) and varies the target accuracy $\varepsilon$. Both methods appear to scale similarly. However Algorithm 4 is over 5× faster, with the gap widening in the high-accuracy regime.

Fig. 2(c) examines the performance impact of the ratio $L_f/\|A\|^2$ in Algorithm 4 and LPALM. We fix 20 problems $\{(M_i, c_i, A_i, b_i)\}$, then rescale each problem $(M_i, c_i, A_i, b_i)$ to $(\chi M_i, \chi c_i, A_i, b_i)$ and solve to $\chi 10^{-6}$ accuracy for some $\chi > 0$[3]. It is worth emphasizing again that the problems, minimizers, and relative accuracy are constant: the only variable is the rescaled ratio. We tested 10 values of $\chi \in [0.1, 100]$. As seen in Fig. 2(c), LPALM and Algorithm 4 are effective in two very

[3]We hold the primal feasibility target constant, only adjusting the tolerance for the primal subgradient norm.

different regimes. For $L_f \ll \|A\|^2$, Algorithm 4 is over $10\times$ faster. However, the methods meet when $L_f \sim 0.03\|A\|^2$, and LPALM significantly overtakes Algorithm 4 in the regime $L_f \geq 0.1\|A\|^2$. These findings suggest that the "rescaling" discussed in the remarks after Corollary 3.5 is more than a theoretical convenience: Algorithm 4 performs significantly better when $L_f \ll \|A\|^2$, even when that requires decreasing $\varepsilon$.

# 7 Concluding Remarks

This paper proposes the Restarted ACG method (Algorithm 2), I-ALM (Algorithm 3), and I-FALM (Algorithm 4). Our improved analysis of all three methods is grounded in a unified IPP perspective, making use of the LOrA and FLOrA frameworks proposed in Section 4. Using the FLOrA framework, we show that Algorithm 2 achieves optimal global complexity for solving (1) in both convex and strongly convex settings, which, to our knowledge, is a novel result in the restarted ACG literature. Similarly, we utilize the LOrA framework to prove that Algorithm 3 achieves near-optimal, non-ergodic complexity for solving (2) with constant regularization, a novel result in the ALM literature to our knowledge. Finally, we combine the analysis of Algorithm 3 with the FLOrA framework to develop an accelerated variant, I-FALM (Algorithm 4), which also achieves near-optimal non-ergodic complexity for solving (2). Both Algorithms 3 and 4 utilize gradient mapping-based termination criteria for the inner ACG solver, which are both efficiently computable and remove the need for the postprocessing used in previous ALM literature [18]. Numerical experiments validate the empirical performance of the proposed algorithms, with Algorithm 4 significantly outperforming competing ALM variants.

Several related questions merit future investigation. First, Restarted ACG attains optimal complexity for strongly convex optimization if the modulus $\mu_f$ is provided. However, in the absence of prior knowledge about $\mu_f$, one must rely on universal methods such as [14, 49], which achieve complexity bounds in terms of $\mu_f$ as good as those obtained when $\mu_f$ is known in advance. Second, Assumption 2(d) plays a crucial role in our analysis throughout Section 3, and it remains an open question whether optimal I-ALM variants can be designed with inexact subroutines without boundedness. It is also of interest to design an algorithm that does not require an estimate of $R_\Lambda$ as input, since an estimate may not be available a priori. Third, another related pursuit would be to obtain (near)-optimal complexities for Algorithm 4 without primal-dual perturbations (i.e., $\gamma_d = 0$ and $\gamma_p = 0$), which may remove the explicit need for an $R_\Lambda$ estimate.

# References


[1] T. Alamo, P. Krupa, and D. Limon. Gradient based restart FISTA. In *58th IEEE Conference on Decision and Control (CDC)*, pages 3936–3941. IEEE, 2019.

[2] T. Alamo, P. Krupa, and D. Limon. Restart of accelerated first order methods with linear convergence under a quadratic functional growth condition. *IEEE Transactions on Automatic Control*, 67(10):5200–5214, 2022.

[3] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.

[4] C. Bao, L. Chen, J. Li, and Z. Shen. Accelerated gradient methods with gradient restart: Global linear convergence. *arXiv preprint arXiv:2401.07672*, 2024.

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

[6] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical Optimization: Theoretical and Practical Aspects*. Springer, 2006.

[7] G. Braun, A. Carderera, C. W. Combettes, H. Hassani, A. Karbasi, A. Mokhtari, and S. Pokutta. *Conditional Gradient Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2025.

[8] S. Bubeck, Q. Jiang, Y. Lee, T. Li, and A. Sidford. Near-optimal method for highly smooth convex optimization. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 492–507. PMLR, 2019.

[9] K. Deng, R. Wang, Z. Zhu, J. Zhang, and Z. Wen. The augmented Lagrangian methods: Overview and recent advances. *arXiv preprint arXiv:2510.16827*, 2025.

[10] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002.

[11] M. Frank, P. Wolfe, et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[12] G. Frison and M. Diehl. HPIPM: A high-performance quadratic programming framework for model predictive control. *IFAC-PapersOnLine*, 53(2):6563–6569, 2020.

[13] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, and C. Uribe. Optimal tensor methods in smooth convex and uniformly convex optimization. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 1374–1391. PMLR, 2019.

[14] V. Guigues, J. Liang, and R. D. C. Monteiro. Universal subgradient and proximal bundle methods for convex and strongly convex hybrid composite optimization. *Journal of Optimization Theory and Applications*, 208(3):112, 2026.

[15] M. R. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, 1969.

[16] M. Ito and M. Fukuda. Nearly optimal first-order methods for convex optimization under gradient norm measure: An adaptive regularization approach. *Journal of Optimization Theory and Applications*, 188(3):770–804, 2021.

[17] B. Jiang, H. Wang, and S. Zhang. An optimal high-order tensor method for convex optimization. *Mathematics of Operations Research*, 46(4):1390–1412, 2021.

[18] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Mathematical Programming*, 155(1-2):511–547, 2016.

[19] C. Lemaréchal. An extension of Davidon methods to non differentiable problems. In *Nondifferentiable optimization*, pages 95–109. Springer, 1975.

[20] C. Lemaréchal. Nonsmooth optimization and descent methods. *IIASA*, 1978.

[21] H. Li and Z. Lin. Accelerated alternating direction method of multipliers: An optimal O(1 / K) nonergodic analysis. *Journal of Scientific Computing*, 79(2):671–699, 2019.

[22] J. Liang. Primal-dual proximal bundle and conditional gradient methods for convex problems. *Mathematical Programming*, pages 1–48, 2025.

[23] J. Liang and R. D. C. Monteiro. A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes. *SIAM Journal on Optimization*, 31(4):2955–2986, 2021.

[24] J. Liang and R. D. C. Monteiro. A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems. *Mathematics of Operations Research*, 49(2):832–855, 2024.

[25] Y. F. Liu, X. Liu, and S. Ma. On the nonergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming. *Mathematics of Operations Research*, 44(2):632–650, 2019.

[26] Z. Lu and Z. Zhou. Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. *SIAM Journal on Optimization*, 33(2):1159–1190, 2023.

[27] M. Marques Alves. Variants of the A-HPE and large-step A-HPE algorithms for strongly convex problems with applications to accelerated high-order tensor methods. *Optimization Methods and Software*, 37(6):2021–2051, 2022.

[28] J. G. Melo, R. D. C. Monteiro, and H. Wang. A proximal augmented Lagrangian method for linearly constrained nonconvex composite optimization problems. *Journal of Optimization Theory and Applications*, 202(1):388–420, 2024.

[29] R. Mifflin. A modification and an extension of Lemaréchal's algorithm for nonsmooth minimization. In *Nondifferential and variational techniques in optimization*, pages 77–90. Springer, 1982.

[30] R. D. C. Monteiro, C. Ortiz, and B. F. Svaiter. An adaptive accelerated first-order method for convex optimization. *Computational Optimization and Applications*, 64:31–73, 2016.

[31] R. D. C. Monteiro and B. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787, 2010.

[32] R. D. C. Monteiro and B. F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

[33] R. D. C. Monteiro and B. F. Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, 23(1):475–507, January 2013.

[34] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.

[35] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[36] Y. Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, 2018.

[37] B. O'Donoghue and E. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.

[38] Y. Ouyang, Y. Chen, G. Lan, and E. Pasiliao. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 8(1):644–681, 2015.

[39] Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1/2):1–35, 2021.

[40] A. Patrascu, I. Necoara, and Quoc Tran-Dinh. Adaptive inexact fast augmented Lagrangian methods for constrained convex optimization. *Optimization Letters*, 11(3):609–626, 2017.

[41] M. J. Powell. A method for nonlinear constraints in minimization problems. *Optimization*, pages 283–298, 1969.

[42] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1(2):97–116, 1976.

[43] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[44] S. Sabach and M. Teboulle. Faster Lagrangian-based methods in convex optimization. *SIAM Journal on Optimization*, 32(1):204–227, 2022.

[45] M. V. Solodov and B. F. Svaiter. A hybrid approximate extragradient – proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7(4):323–345, 1999.

[46] M. V. Solodov and B. F. Svaiter. A hybrid projection-proximal point algorithm. *Journal of Convex Analysis*, 6(1):59–70, 1999.

[47] M. V. Solodov and B. F. Svaiter. An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. *Mathematics of Operations Research*, 25(2):214–230, 2000.

[48] W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[49] A. Susanna and R. D. C. Monteiro. Efficient parameter-free restarted accelerated gradient methods for convex and strongly convex optimization. *Journal of Optimization Theory and Applications*, 206(2):52, 2025.

[50] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[51] P. Wolfe. A method of conjugate subgradients for minimizing nondifferentiable functions. In *Nondifferentiable optimization*, pages 145–173. Springer, 1975.

[52] Y. Xu. Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming. *SIAM Journal on Optimization*, 27(3):1459–1484, 2017.

[53] Y. Xu. Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Mathematical Programming*, 185(1):199–244, 2021.

# A Details of Numerical Experiments

All experiments are run on a workstation desktop running Ubuntu 24.04.3 LTS with an Intel i9-13900k CPU and 64 GB of RAM. Proximal operator implementations are taken from the ProximalOperators.jl[4] package.

## A.1 Restarted ACG Experimental Details

Recall that our problem of interest is the sparse linear regression/LASSO problem (98). For testing we set $n = 1000$, $m = 500$, and $\gamma = 1/2$. $A$ is set to 20% density, with nonzero entries generated IID normal. The vector $b$ is randomly generated with IID uniform entries over $[0, 1]$. We start each solver from the origin $x_0 = 0$ and use $L_f = \|A\|^2$. After some brief parameter tuning, we set $\lambda = 0.2$ in Algorithm 2. The function value and number of restart steps are logged on every restart for each algorithm. Note the location of data points along the x-axis of Fig. 1 is therefore non-uniform, since the number of steps between each restart differ for each algorithm.

## A.2 I-ALM Experimental Details

Recall that the problem class used for I-ALM testing is the linearly constrained quadratic program

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top M x + c^\top x \\ &\text{s.t. } Ax = b \\ &\quad x_\ell \leq x_i \leq x_u \text{ for all } i \in \{1, \ldots, n\}. \end{aligned}$$

Fixing $n, m, r \in \mathbb{N}$ with $n \geq \max\{r, m\}$ and $\delta \in [0, 1]$, we generate problem structures using the following procedure:

- $\tilde{M} = RR^T$ with $R \in \mathbb{R}^{n \times r}$, $R_{ij} \sim \mathcal{N}(0, 1)$. We then rescale $M_{ij} = \tilde{M}_{ij}/\|M\|$ to ensure that $L_f = \|M\| = 1$,
- $c \in \mathbb{R}^n$, $c_i \sim \mathcal{N}(0, 1)$,
- $A \in \mathbb{R}^{m \times n}$, $A_{ij} \sim \text{Bern}(\delta) \cdot \mathcal{N}(0, 1)$,
- $b \in \mathbb{R}^m$, $b_i \sim \mathcal{N}(0, 1)$,
- $x_\ell = -10$, $x_u = 10$.

$M$ is therefore an $n \times n$ matrix with rank $r$, $c$ and $b$ are entry-wise normally-distributed vectors, and $A$ is a normally distributed matrix with density $\delta = 0.1$. We them compute the diameter $D$ as $D = \sqrt{n}(x_\ell - x_u)$. We estimate $\hat{R}_\Lambda = 1000$ for all problems, which worked as a sufficient upper bound in practice. This procedure was only used for testing. In general, one could use a "guess-and-check" procedure as proposed in [30] which only adds $\mathcal{O}(\log \hat{R}_{\tilde{\Lambda}})$ complexity. For "Lu and Zhou", we set $\varepsilon_k = \varepsilon_0 \alpha^k$ with $\alpha = 0.7$, $\varepsilon_0 = 0.1$, and $\rho_0 = 10$ after preliminary tuning. For LPALM we set $\rho = \max\{\sqrt{L_f}/\|A\|, L_f/\|A\|^2\}$, which was a performant heuristic in our limited numerical testing. We set $\alpha = 0.7$, $\varepsilon_0 = 100$ for Algorithm 3 and $\alpha = 0.85$, $\varepsilon_0 = \rho^{-1}$ for Algorithm (4).

[4] `https://github.com/JuliaFirstOrder/ProximalOperators.jl`

# B Technical Results

The gradient mapping (defined in (13)) serves a critical role in Subsections 5.2 and 5.3, as well as in the numerical experiments in Section 6. The following lemma provides several technical results for the gradient mapping of a convex composite function.

**Lemma B.1.** *Consider problem* (1)*, which we assume satisfies Assumption 1. Additionally assume that* $\operatorname{dom} h$ *is bounded with diameter* $D$*. Given* $\eta \leq L_f^{-1}$ *and* $\tilde{x} \in \operatorname{dom} h$*, define the gradient mapping* $\mathcal{G}_\phi^\eta(\tilde{x})$ *as in* (13) *and set*

$$x^+ = \tilde{x} - \eta \mathcal{G}_\phi^\eta(\tilde{x}). \tag{99}$$

*Then, the following statements hold:*

a) *denoting* $\phi_* = \min_{x \in \operatorname{dom} h} \phi(x)$*, we have*

$$\phi(x^+) - \phi_* \leq D\|\mathcal{G}_\phi^\eta(\tilde{x})\| - \frac{\eta}{2}\|\mathcal{G}_\phi^\eta(\tilde{x})\|^2;$$

b) *if* $\|\mathcal{G}_\phi^\eta(\tilde{x})\| \leq \varepsilon$*, then there exists a subgradient* $v \in \partial\phi(x^+)$ *satisfying* $\|v\| \leq 2\varepsilon$*;*

c) *given* $\gamma > 0$ *and* $\bar{x} \in \operatorname{dom} h$*, define* $\phi_\gamma(x) = \phi(x) + \gamma\|x - \bar{x}\|^2/2$*. Suppose that* $\gamma \leq \varepsilon/(2D)$ *and* $\|\mathcal{G}_{\phi_\gamma}^\eta(\tilde{x})\| \leq \varepsilon/2$*, where the proximal mapping in* $\mathcal{G}_{\phi_\gamma}^\eta(\cdot)$ *is still with respect to* $h$*, then* $\|\mathcal{G}_\phi^\eta(\tilde{x})\| \leq \varepsilon$*.*

**Proof**: a) Applying Lemma 2.3 of [5] and using the definition of $x^+$ in (99), we have for every $y \in \operatorname{dom} h$,

$$\begin{aligned}\phi(y) - \phi(x^+) &\geq \frac{1}{2\eta}\|x^+ - \tilde{x}\|^2 + \eta^{-1}\langle \tilde{x} - y, x^+ - \tilde{x}\rangle \overset{(99)}{=} \frac{\eta}{2}\|\mathcal{G}_\phi^\eta(\tilde{x})\|^2 - \langle \tilde{x} - y, \mathcal{G}_\phi^\eta(\tilde{x})\rangle \\ &\geq \frac{\eta}{2}\|\mathcal{G}_\phi^\eta(\tilde{x})\|^2 - \|\tilde{x} - y\|\|\mathcal{G}_\phi^\eta(\tilde{x})\| \geq \frac{\eta}{2}\|\mathcal{G}_\phi^\eta(\tilde{x})\|^2 - D\|\mathcal{G}_\phi^\eta(\tilde{x})\|,\end{aligned}$$

where the second inequality is due to the Cauchy-Schwarz inequality and the last inequality follows from the boundedness of $\operatorname{dom} h$. The statement follows by taking $y = x_*$ for any $x_* \in \{x \in \operatorname{dom} h : \phi(x) = \phi_*\}$.

b) In view of (13), the definition of $x^+$ in (99) can be rewritten as

$$x^+ = \operatorname{prox}_{\eta h}(\tilde{x} - \eta \nabla f(\tilde{x})),$$

whose optimality condition yields that

$$0 \in \frac{x^+ - \tilde{x}}{\eta} + \nabla f(\tilde{x}) + \partial h(x^+).$$

Rearranging terms and adding $\nabla f(x^+)$ to both sides, we have

$$v := \frac{\tilde{x} - x^+}{\eta} - \nabla f(\tilde{x}) + \nabla f(x^+) \in \partial h(x^+) + \nabla f(x^+) = \partial\phi(x^+).$$

Using the triangle inequality and the smoothness of $f$, we have

$$\|v\| \leq \frac{1}{\eta}\|\tilde{x} - x^+\| + L_f\|x^+ - \tilde{x}\| \leq \frac{2}{\eta}\|\tilde{x} - x^+\| \overset{(99)}{=} 2\|\mathcal{G}_\phi^\eta(\tilde{x})\| \leq 2\varepsilon,$$

where the second inequality is due to the fact that $L_f \leq 1/\eta$. Hence, we prove the statement.

c) It follows from Lemma 3.1(iii) of [16] that for any $\gamma > 0$,

$$\|\mathcal{G}_{\phi}^{\eta}(\tilde{x}) - \mathcal{G}_{\phi_\gamma}^{\eta}(\tilde{x})\| \leq \gamma \|\tilde{x} - \bar{x}\|. \tag{100}$$

Using the above inequality and the triangle inequality, we have

$$\|\mathcal{G}_{\phi}^{\eta}(\tilde{x})\| \leq \|\mathcal{G}_{\phi_\gamma}^{\eta}(\tilde{x})\| + \|\mathcal{G}_{\phi_\gamma}^{\eta}(\tilde{x}) - \mathcal{G}_{\phi}^{\eta}(\tilde{x})\| \overset{(100)}{\leq} \|\mathcal{G}_{\phi_\gamma}^{\eta}(\tilde{x})\| + \gamma\|\tilde{x} - \bar{x}\| \leq \|\mathcal{G}_{\phi_\gamma}^{\eta}(\tilde{x})\| + \gamma D \leq \varepsilon,$$

where the third inequality follows from boundedness, and the final inequality follows from the assumptions on $\|\mathcal{G}_{\phi_\gamma}^{\eta}(\tilde{x})\|$ and $\gamma$. ∎

The following result connects the gradient mapping of the augmented Lagrangian function $\mathcal{L}_\rho$ to the subdifferential of the Lagrangian $\mathcal{L}$. The lemma is used to prove Propositions 5.6 and 5.10.

**Proposition B.2.** *Consider problem* (2)*, which we assume satisfies Assumption 2. Given* $\lambda \in \mathbb{R}^m$ *and* $\rho > 0$*, let* $\mathcal{L}_\rho(\cdot, \lambda)$ *be the augmented Lagrangian in* (8)*. Define* $M_\rho$ *as in* (32)*, let* $\eta \leq M_\rho^{-1}$*, and suppose* $\tilde{x}$ *is a point satisfying* $\|\mathcal{G}_{\mathcal{L}_\rho(\cdot,\lambda)}^{\eta}(\tilde{x})\| \leq \varepsilon/2$*. Set*

$$x^+ = \tilde{x} - M_\rho^{-1}\mathcal{G}_{\mathcal{L}_\rho(\cdot,\lambda)}^{\eta}(\tilde{x}), \quad \lambda^+ = \lambda + \rho(Ax^+ - b). \tag{101}$$

*Then, there exists a subgradient* $v \in \partial\mathcal{L}(\cdot, \lambda^+)(x^+)$ *satisfying* $\|v\| \leq \varepsilon$*.*

**Proof**: By Lemma B.1(b) applied to $\mathcal{L}_\rho(\cdot, \lambda)$ with gradient mapping stepsize $\eta$, there exists a $v \in \partial\mathcal{L}_\rho(\cdot, \lambda)(x^+)$ satisfying $\|v\| \leq \varepsilon$. It follows from the definition of $\mathcal{L}_\rho(\cdot, \lambda)$ in (8) and subdifferential calculus that

$$\begin{aligned}\partial\mathcal{L}_\rho(\cdot, \lambda)(x^+) &= \partial\left(\mathcal{L}(\cdot, \lambda) + \frac{\rho}{2}\|A \cdot -b\|^2\right)(x^+) = \partial\mathcal{L}(\cdot, \lambda)(x^+) + \rho A^\top(Ax^+ - b) \\ &= \partial(\mathcal{L}(\cdot, \lambda) + \langle \rho(Ax^+ - b), A \cdot -b\rangle)(x^+) \overset{(3),(101)}{=} \partial\mathcal{L}(\cdot, \lambda^+)(x^+),\end{aligned}$$

where the last identity follows from (101) and the definition of $\mathcal{L}(\cdot, \lambda)$ in (3). Therefore, we prove $v \in \partial\mathcal{L}(\cdot, \lambda^+)(x^+)$ and thus conclude the proof. ∎

We can show that an $\varepsilon$-primal-dual solution to (2) in the sense of (4) implies an $\mathcal{O}(\varepsilon)$ bound on the absolute primal gap $|\phi(x) - \hat{\phi}_*|$. The result has been used in prior works [26], though we repeat the proof for completeness.

**Lemma B.3.** *Suppose* $(x, \lambda)$ *is an* $\varepsilon$*-primal-dual solution to* (2) *in the sense of* (4)*. Then the absolute value of the primal gap is bounded by*

$$|\phi(x) - \hat{\phi}_*| \leq \varepsilon \max\{\|\lambda_*\|, \|\lambda\| + D\}, \tag{102}$$

*where* $\lambda_* \in \Lambda_* = \{\lambda : d(\lambda) = d_*\}$ *is an optimal dual solution to* (3)*.*

**Proof**: We start by proving

$$\phi(x) - d(\lambda) \leq \varepsilon(\|\lambda\| + D). \tag{103}$$

Since $(x, \lambda)$ is an $\varepsilon$-primal-dual solution to (2),

$$v \in \partial\mathcal{L}(\cdot, \lambda)(x), \quad \|v\| \leq \varepsilon, \quad \|Ax - b\| \leq \varepsilon, \tag{104}$$

for some $v \in \mathbb{R}^n$. Define $u(\lambda) = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \mathcal{L}(x, \lambda)$. Then by the definition of the subdifferential $\partial\mathcal{L}(\cdot, \lambda)(x)$ we have

$$-d(\lambda) \overset{(3)}{=} -\mathcal{L}(u(\lambda), \lambda) \leq -\mathcal{L}(x, \lambda) - \langle v, u(\lambda) - x\rangle \overset{(3)}{=} -\phi(x) - \langle \lambda, Ax - b\rangle - \langle v, u(\lambda) - x\rangle.$$

Rearranging and using the Cauchy-Schwarz inequality gives

$$\begin{aligned}\phi(x) - d(\lambda_r) &\leq -\langle \lambda, Ax - b\rangle - \langle v, u(\lambda) - x\rangle \\ &\leq \|\lambda\|\|Ax - b\| + \|v\|\|u(\lambda) - x\| \overset{(104)}{\leq} \varepsilon\|\lambda\| + \varepsilon D,\end{aligned}$$

where the last inequality follows from (104) and Assumption 2(d).

We now prove (102). Clearly, the upper bound

$$\phi(x) - \hat{\phi}_* \leq \varepsilon(\|\lambda\| + D)$$

follows from (103) and strong duality $d(\lambda) \leq d_* = \hat{\phi}_*$ by (3). Let $x_* \in \{x \in \operatorname{dom} h : \phi(x) = \hat{\phi}_*,\ Ax = b\}$ be an optimizer of (2). Note that since $(x_*, \lambda_*)$ is a saddle-point of $\mathcal{L}$ in (3), we have

$$\begin{aligned}0 \leq \mathcal{L}(x, \lambda_*) - \mathcal{L}(x_*, \lambda_*) &= \phi(x) + \langle \lambda_*, Ax - b\rangle - \hat{\phi}_* - \langle \lambda_*, Ax_* - b\rangle \\ &= \phi(x) + \langle \lambda_*, Ax - b\rangle - \hat{\phi}_*.\end{aligned}$$

where the last equality follows by the feasibility of $x_*$. Rearranging, we have

$$\phi(x) - \hat{\phi}_* \geq -\langle \lambda_*, Ax - b\rangle \geq -\|\lambda_*\|\|Ax - b\| \overset{(104)}{\geq} -\|\lambda_*\|\varepsilon,$$

where the second inequality follows from the Cauchy-Schwarz inequality and the final inequality follows from the assumption that $(x, \lambda)$ is an $\varepsilon$-primal-dual solution to (2). The inequality (102) follows from combining the upper and lower bounds. ∎

# C Analysis of Frameworks in Section 4

This section develops the analysis of LOrA and FLOrA frameworks introduced in Section 4 and finally proves the two main results on sub-optimality guarantees, namely Theorems 4.4 and 4.5.

## C.1 LOrA Analysis

We begin the analysis of LOrA by providing the proof of Proposition 4.1. For simplicity of notation, we omit the superscripts $\cdot^{\mathrm{L}}$ and subscripts $\cdot_{\mathrm{L}}$ in all proofs.

**Proof of Proposition 4.1:** We first note from (52) that

$$\hat{u}_{k+1} \in \partial\Gamma_k(x_{k+1}). \tag{105}$$

Using (50), (105), and the fact that $\Gamma_k$ is $\lambda^{-1}$-strongly convex, we have

$$\Phi(\hat{x}_*) + \frac{1}{2\lambda}\|\hat{x}_* - x_k\|^2 \overset{(50)}{\geq} \Gamma_k(\hat{x}_*) \overset{(105)}{\geq} \Gamma_k(x_{k+1}) + \langle \hat{u}_{k+1}, \hat{x}_* - x_{k+1}\rangle + \frac{1}{2\lambda}\|\hat{x}_* - x_{k+1}\|^2.$$

Rearranging the terms and adding $\lambda\|\hat{u}_{k+1}\|^2/2$ to both sides, we have

$$\begin{aligned}\frac{\lambda}{2}\|\hat{u}_{k+1}\|^2 - \Gamma_k(x_{k+1}) + \Phi(\hat{x}_*) + \frac{1}{2\lambda}\|\hat{x}_* - x_k\|^2 &\geq \frac{\lambda}{2}\|\hat{u}_{k+1}\|^2 + \langle \hat{u}_{k+1}, \hat{x}_* - x_{k+1}\rangle + \frac{1}{2\lambda}\|\hat{x}_* - x_{k+1}\|^2 \\ &= \frac{1}{2\lambda}\|\lambda\hat{u}_{k+1} + \hat{x}_* - x_{k+1}\|^2 \geq 0,\end{aligned}$$

and hence

$$\frac{\lambda}{2}\|\hat{u}_{k+1}\|^2 - \Gamma_k(x_{k+1}) \geq -\Phi(\hat{x}_*) - \frac{1}{2\lambda}\|\hat{x}_* - x_k\|^2. \tag{106}$$

Combining the above inequality with (51) yields

$$\begin{aligned}\frac{\sigma}{2\lambda}\|y_{k+1} - x_k\|^2 + \delta_k &\overset{(51)}{\geq} \frac{\lambda}{2}\|\hat{u}_{k+1}\|^2 + \Phi(y_{k+1}) + \frac{1}{2\lambda}\|y_{k+1} - x_k\|^2 - \Gamma_k(x_{k+1}) \\ &\overset{(106)}{\geq} \Phi(y_{k+1}) + \frac{1}{2\lambda}\|y_{k+1} - x_k\|^2 - \Phi(\hat{x}_*) - \frac{1}{2\lambda}\|\hat{x}_* - x_k\|^2,\end{aligned}$$

proving the claim. ∎

We next present a technical lemma that is useful in the analysis of LOrA. The first statement is a single-step bound on the primal gap, while the second statement provides a uniform upper bound for the iterate distance. Note that if $\delta_k^{\mathrm{L}} = 0$ for all $k \geq 0$, then the sequence $\|x_k^{\mathrm{L}} - x_*\|$ is non-increasing. However, uniformly bounding the distance with $\delta_k^{\mathrm{L}} > 0$ requires summability of the absolute error terms.

**Lemma C.1.** *Let $X_*$ be the set of optimal solutions to* (49)*. Define $x_* = \operatorname{argmin}\{\|x_0^{\mathrm{L}} - x\| : x \in X_*\}$ and suppose $\mathcal{A}_k^{\mathrm{L}} = \infty$. Then, for every $k \geq 0$, we have*

$$2\lambda_{\mathrm{L}}[\Phi(y_{k+1}^{\mathrm{L}}) - \Phi(x_*)] + (1 - \sigma_{\mathrm{L}})\|y_{k+1}^{\mathrm{L}} - x_k^{\mathrm{L}}\|^2 \leq \|x_k^{\mathrm{L}} - x_*\|^2 - \|x_{k+1}^{\mathrm{L}} - x_*\|^2 + 2\lambda\delta_k^{\mathrm{L}}. \tag{107}$$

*Moreover, if $\{\delta_k^{\mathrm{L}}\}$ is summable with $C_\delta := \sum_{i=0}^{\infty}\delta_k^{\mathrm{L}} < \infty$, we have for every $k \geq 0$,*

$$\|x_k^{\mathrm{L}} - x_*\| \leq R_0^{\mathrm{L}} + \sqrt{2\lambda_{\mathrm{L}} C_\delta}, \tag{108}$$

*where $R_0^{\mathrm{L}} = \|x_0^{\mathrm{L}} - x_*\|$.*

**Proof**: Using (51) and the fact that the objective in (52) is $(\mathcal{A}_k^{-1} + \lambda^{-1})$-strongly convex, we have for every $v \in \mathbb{R}^n$,

$$\begin{aligned}\Gamma_k(v) + \frac{1}{2\mathcal{A}_k}\|v - x_k\|^2 - \frac{\lambda + \mathcal{A}_k}{2\lambda\mathcal{A}_k}\|v - x_{k+1}\|^2 + \delta_k &\overset{(52)}{\geq} \Gamma_k(x_{k+1}) + \frac{1}{2\mathcal{A}_k}\|x_{k+1} - x_k\|^2 + \delta_k \\ &\overset{(51)}{\geq} \frac{1}{2\lambda}\|\lambda\hat{u}_{k+1}\|^2 + \Phi(y_{k+1}) + \frac{1-\sigma}{2\lambda}\|y_{k+1} - x_k\|^2.\end{aligned}$$

The above inequality together with (50) implies that

$$\Phi(v) + \frac{\lambda + \mathcal{A}_k}{2\mathcal{A}_k\lambda}\|v - x_k\|^2 - \frac{\lambda + \mathcal{A}_k}{2\mathcal{A}_k\lambda}\|v - x_{k+1}\|^2 + \delta_k \geq \frac{1}{2\lambda}\|\lambda\hat{u}_{k+1}\|^2 + \Phi(y_{k+1}) + \frac{1-\sigma}{2\lambda}\|y_{k+1} - x_k\|^2.$$

Hence, (107) immediately follows by taking $v = x_*$ and $\mathcal{A}_k = \infty$ and rearranging the terms.

Rearranging (107) and discarding non-negative terms, we have for every $k \geq 0$,

$$\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 + 2\lambda\delta_k.$$

Summing both sides from 0 to $k-1$ gives

$$\|x_k - x_*\|^2 \leq \|x_0 - x_*\|^2 + 2\lambda\sum_{i=0}^{k-1}\delta_0 \leq 2\lambda C_\delta,$$

which proves (108) using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$. ∎

We are now ready to prove Theorem 4.4, which follows directly from the single-step claim of Lemma C.1.

**Proof of Theorem 4.4:** Summing (107) from $k = 0$ to $k - 1$ we have

$$2\lambda \sum_{i=0}^{k-1} [\Phi(y_{i+1}) - \Phi(x_*)] + (1-\sigma) \sum_{i=0}^{k-1} \|y_{i+1} - x_i\|^2 \leq \sum_{i=0}^{k-1} \left( \|x_i - x_*\|^2 - \|x_{i+1} - x_*\|^2 \right) + 2\lambda_L \sum_{i=0}^{k-1} \delta_i.$$

Using the definitions of $R_0$ and $\bar{\delta}_k$ given in the theorem, we obtain

$$2\lambda k \min_{1 \leq i \leq k} [\Phi(y_i) - \Phi(x_*)] + (1-\sigma) k \min_{1 \leq i \leq k} \|y_i - x_{i-1}\|^2 \leq \|x_0 - x_*\|^2 + 2\lambda k \bar{\delta}_k,$$

and hence conclude the claims. ∎

## C.2 FLOrA Analysis

This subsection is devoted to the analysis of FLOrA, which is introduced in Subsection 4.2. For simplicity of notation, we omit the superscripts $\cdot^{\mathrm{F}}$ and subscripts $\cdot_{\mathrm{F}}$ in all proofs.

Many of the following results are analogous to those in the analysis of accelerated first-order methods, however our inclusion of the absolute error sequence $\{\delta_k^{\mathrm{F}}\}$ in Algorithm 6 requires modification of some statements and/or proofs.

**Lemma C.2.** *For every $k \geq 0$, the following statements hold:*

a) $b_k^2 = \tau_k \lambda_{\mathrm{F}} B_{k+1}$;

b) $\tau_k = 1 + \mu_{\mathrm{F}} B_k$;

c)
$$B_{k+1} \geq \lambda_{\mathrm{F}} \max \left\{ \frac{(k+1)^2}{4}, \left( 1 + \frac{\sqrt{\lambda_{\mathrm{F}} \mu_{\mathrm{F}}}}{2} \right)^{2k} \right\};$$

d) *recall $C_{\mathrm{F}} = \sum_{i=0}^{\infty} B_{i+1} \alpha_{\mathrm{F}}^i$ defined in Theorem 4.5, then $C_{\mathrm{F}} < \infty$. Furthermore, denoting $\beta_{\mathrm{F}} = \sqrt{\alpha_{\mathrm{F}}}(1 + \sqrt{\lambda_{\mathrm{F}} \mu_{\mathrm{F}}}) < 1$, we have $C_{\mathrm{F}} \leq \lambda_{\mathrm{F}} / (1 - \beta_{\mathrm{F}})^4$;*

e) *define the sequence $\{\Delta_k^{\mathrm{F}}\}$ as*

$$\Delta_{-1}^{\mathrm{F}} = 0, \quad \Delta_k^{\mathrm{F}} = \frac{B_k}{B_{k+1}} \Delta_{k-1}^{\mathrm{F}} + \delta_k^{\mathrm{F}}, \tag{109}$$

*then we have*

$$\Delta_k^{\mathrm{F}} = \frac{\delta_0^{\mathrm{F}}}{B_{k+1}} \sum_{i=0}^{k} B_{i+1} (\alpha_{\mathrm{F}})^i \leq \frac{\delta_0^{\mathrm{F}} C_{\mathrm{F}}}{B_{k+1}}. \tag{110}$$

**Proof**: a) It is easy to verify that $b_k$ in (61) is the root of equation $b_k^2 - \lambda \tau_k b_k - \lambda \tau_k B_k = 0$, which is equivalent to statement a) in view of the second identity in (61).

b) This statement immediately follows from the second and last equations in (61) and $B_0 = 0$.

c) This statement can be easily shown in a way similar to the proof of Proposition 1(c) of [30] and hence we omit the proof.

d) Using (61) and the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$, we have

$$\begin{aligned} B_{k+1} = B_k + b_k &\overset{(61)}{\le} B_k + \lambda\tau_k + \sqrt{\lambda\tau_k B_k} \le (\sqrt{B_k} + \sqrt{\lambda\tau_k})^2 \\ &= (\sqrt{B_k} + \sqrt{\lambda(1+\mu B_k)})^2 \le [(1+\sqrt{\lambda\mu})\sqrt{B_k} + \sqrt{\lambda}]^2, \end{aligned}$$

where the last identity is due to statement b). We thus have the recursion

$$\sqrt{B_{k+1}} \le (1+\sqrt{\lambda\mu})\sqrt{B_k} + \sqrt{\lambda}.$$

Note that $\beta := \sqrt{\alpha}(1+\sqrt{\lambda\mu}) < 1$ by the requirement $\alpha < (1+\sqrt{\lambda\mu})^{-2}$ from the initialization of Algorithm 6. Then, we obtain for all $k \ge 0$

$$\sqrt{\alpha^{k+1} B_{k+2}} \le \beta \sqrt{\alpha^k B_{k+1}} + \sqrt{\alpha}^{k+1}\sqrt{\lambda}.$$

Unrolling with initial element $\alpha^0 B_1 = \lambda$, we obtain the upper bound

$$\sqrt{\alpha^k B_{k+1}} \le \sqrt{\lambda} \sum_{i=0}^{k} \beta^{k-i} \sqrt{\alpha}^i.$$

Then, summing from 0 to $\infty$, we obtain

$$\sum_{k=0}^{\infty} \sqrt{\alpha^k B_{k+1}} \le \sqrt{\lambda} \sum_{k=0}^{\infty} \sum_{i=0}^{k} \beta^{k-i} \sqrt{\alpha}^i = \sqrt{\lambda} \sum_{i=0}^{\infty} \sqrt{\alpha}^i \sum_{k=i}^{\infty} \beta^{k-i} = \frac{\sqrt{\lambda}}{1-\beta} \sum_{i=0}^{\infty} \sqrt{\alpha}^i = \frac{\sqrt{\lambda}}{(1-\sqrt{\alpha})(1-\beta)}.$$

Hence, we have

$$\sum_{k=0}^{\infty} \alpha^k B_{k+1} \le \left( \sum_{k=0}^{\infty} \sqrt{\alpha^k B_{k+1}} \right)^2 \le \frac{\lambda}{(1-\sqrt{\alpha})^2 (1-\beta)^2} \le \frac{\lambda}{(1-\beta)^4},$$

where the last inequality follows from $\beta \ge \sqrt{\alpha}$.

e) It follows from (109) and $\delta_k = \delta_0 \alpha^k$ (see Step **1** of Algorithm 6) that

$$B_{k+1}\Delta_k = \sum_{i=0}^{k} B_{i+1}\delta_i = \delta_0 \sum_{i=0}^{k} B_{i+1}\alpha^i.$$

Hence, we prove (110) in view of statement d). ∎

**Lemma C.3.** *For every $k \ge 0$, define*

$$\theta_{k+1}(x) = \Gamma_k^{\mathrm{F}}(z_{k+1}^{\mathrm{F}}) - \frac{1}{2\lambda_{\mathrm{F}}} \|z_{k+1}^{\mathrm{F}} - \tilde{x}_k^{\mathrm{F}}\|^2 + \langle u_{k+1}^{\mathrm{F}}, x - z_{k+1}^{\mathrm{F}} \rangle + \frac{\mu_{\mathrm{F}}}{2} \|x - z_{k+1}^{\mathrm{F}}\|^2, \tag{111}$$

$$\Theta_{k+1}(x) = \frac{B_k \Theta_k(x) + b_k \theta_{k+1}(x)}{B_{k+1}} + \delta_k^{\mathrm{F}}, \tag{112}$$

*with $\Theta_0 \equiv 0$. Then, for every $k \ge 0$, the following statements hold:*

a) *$\theta_{k+1}$ and $\Theta_{k+1}$ are $\mu_{\mathrm{F}}$-strongly convex quadratic functions;*

b) $x_k^{\mathrm{F}} = \operatorname*{argmin}_{u \in \mathbb{R}^n} \{B_k \Theta_k(u) + \|u - x_0^{\mathrm{F}}\|^2/2\}$;

c) *for all* $x \in \operatorname{dom}\Phi$, $\theta_{k+1}(x) \le \Phi(x)$ *and* $\Theta_{k+1}(x) \le \Phi(x) + \Delta_k^{\mathrm{F}}$.

**Proof**: a) The statement simply follows from the definitions of $\theta_{k+1}$ and $\Theta_{k+1}$ in (111) and (112), respectively.

b) We prove the statement by induction. Since $B_0 = 0$ and $\Theta_0 \equiv 0$, we trivially have the base case $k = 0$. We assume the statement holds for some $k \ge 0$. Since $\Theta_k$ is a quadratic function with $\nabla^2\Theta_k = \mu I$, Taylor expansion around the minimizer $x_k$ yields the following equality for all $x \in \mathbb{R}^n$,

$$B_k\Theta_k(x) + \frac{1}{2}\|x_0 - x\|^2 = B_k\Theta_k(x_k) + \frac{1}{2}\|x_0 - x_k\|^2 + \frac{1 + B_k\mu}{2}\|x_k - x\|^2.$$

Hence, it follows from (112) that

$$B_{k+1}\Theta_{k+1}(x) + \frac{1}{2}\|x_0 - x\|^2 = B_k\Theta_k(x_k) + \frac{1}{2}\|x_0 - x_k\|^2 + \frac{1 + B_k\mu}{2}\|x_k - x\|^2 + b_k\theta_{k+1}(x) + B_{k+1}\delta_k.$$

In view of (111), the stationarity condition of $B_{k+1}\Theta_{k+1}(x) + \|x_0 - x\|^2/2$ is

$$0 = (1 + B_k\mu)(x - x_k) + b_k u_{k+1} + b_k\mu(x - z_{k+1}).$$

It is thus straightforward to verify that $x_{k+1}$ in (66) is the solution to the above equation using Lemma C.2(b). Therefore, we complete the proof by induction and conclude the statement.

c) Noting from (65) that $\hat{u}_{k+1} \in \Gamma_k(z_{k+1})$, which together with the first identity in (66) implies that

$$u_{k+1} \in \partial\left(\Gamma_k(\cdot) - \frac{1}{2\lambda}\|\cdot - \tilde{x}_k\|^2\right).$$

It follows from (63) and the fact that $\Gamma_k(\cdot) - \frac{1}{2\lambda}\|\cdot - \tilde{x}_k\|^2$ is $\mu$-strongly convex that for every $x \in \operatorname{dom}\Phi$,

$$\begin{aligned}\Phi(x) &\overset{(63)}{\ge} \Gamma_k(x) - \frac{1}{2\lambda}\|x - \tilde{x}_k\|^2 \\ &\ge \Gamma_k(z_{k+1}) - \frac{1}{2\lambda}\|z_{k+1} - \tilde{x}_k\|^2 + \langle u_{k+1}, x - z_{k+1}\rangle + \frac{\mu}{2}\|x - z_{k+1}\|^2 \overset{(111)}{=} \theta_{k+1}(x),\end{aligned}$$

where the identity is due to the definition of $\theta_{k+1}$ in (111). We have thus proved the first claim.

Using the definitions of $\Theta_{k+1}$ and $\Delta_k$ in (112) and (109), respectively, we can show by induction that

$$\Theta_{k+1}(x) = \frac{\sum_{i=0}^k b_i\theta_{i+1}(x)}{B_{k+1}} + \frac{B_k}{B_{k+1}}\Delta_{k-1} + \delta_k \le \Phi(x) + \Delta_k,$$

where the inequality follows from the first claim $\theta_{k+1} \le \Phi$. ∎

**Lemma C.4.** *For every* $k \ge 0$, *we have*

$$\begin{aligned}&\min_{x\in\mathbb{R}^n}\left\{\Gamma_k^{\mathrm{F}}(z_{k+1}^{\mathrm{F}}) - \frac{1}{2\lambda_{\mathrm{F}}}\|z_{k+1}^{\mathrm{F}} - \tilde{x}_k^{\mathrm{F}}\|^2 + \langle u_{k+1}^{\mathrm{F}}, x - z_{k+1}^{\mathrm{F}}\rangle + \frac{1}{2\lambda_{\mathrm{F}}}\|x - \tilde{x}_k^{\mathrm{F}}\|^2 + \delta_k^{\mathrm{F}}\right\} \\ \ge&\Phi(\tilde{y}_{k+1}^{\mathrm{F}}) + \frac{1 - \sigma_{\mathrm{F}}}{2\lambda_{\mathrm{F}}}\|\tilde{y}_{k+1}^{\mathrm{F}} - \tilde{x}_k^{\mathrm{F}}\|^2. \qquad (113)\end{aligned}$$

**Proof**: In view of (66), we first observe that the minimizer of the left-hand side of (113) is

$$\hat{x} := \tilde{x}_k - \lambda u_{k+1} \overset{(66)}{=} z_{k+1} - \lambda\hat{u}_{k+1}. \qquad (114)$$

Using the second equation in (65), it is easy to verify that

$$\langle u_{k+1}, \hat{x} - z_{k+1}\rangle \overset{(114)}{=} -\langle \hat{u}_{k+1} + \lambda^{-1}(\tilde{x}_k - z_{k+1}), \lambda\hat{u}_{k+1}\rangle \overset{(65)}{=} -\frac{1}{\lambda}\|\lambda\hat{u}_{k+1}\|^2 - \frac{1}{\mathcal{A}_k}\|\tilde{x}_k - z_{k+1}\|^2. \quad (115)$$

Using the above relations and (65) and (66), we have

$$\begin{aligned}
&\Gamma_k(z_{k+1}) - \frac{1}{2\lambda}\|z_{k+1} - \tilde{x}_k\|^2 + \langle u_{k+1}, \hat{x} - z_{k+1}\rangle + \frac{1}{2\lambda}\|\hat{x} - \tilde{x}_k\|^2 \\
&\overset{(114),(115)}{=} \Gamma_k(z_{k+1}) - \frac{1}{2\lambda}\|z_{k+1} - \tilde{x}_k\|^2 - \frac{1}{\lambda}\|\lambda\hat{u}_{k+1}\|^2 - \frac{1}{\mathcal{A}}\|\tilde{x}_k - z_{k+1}\|^2 + \frac{(\mathcal{A}_k + \lambda)^2}{2\lambda\mathcal{A}_k^2}\|z_{k+1} - \tilde{x}_k\|^2 \\
&\overset{(65)}{=} \Gamma_k(z_{k+1}) - \frac{1}{2\lambda}\|\lambda\hat{u}_{k+1}\|^2.
\end{aligned}$$

It follows from (64) that

$$\delta_k + \Gamma_k(z_{k+1}) - \frac{1}{2\lambda}\|\lambda\hat{u}_{k+1}\|^2 \overset{(64)}{\geq} \Phi(\tilde{y}_{k+1}) + \frac{1-\sigma}{2\lambda}\|\tilde{y}_{k+1} - \tilde{x}_k\|^2.$$

Finally, (113) immediately follows from combining the above two relations. ∎

**Lemma C.5.** *Let $X_*$ be the set of optimal solutions to* (49)*. Define $R_0^{\mathrm{F}} := \|x_0^{\mathrm{F}} - x_*\| = \min\{\|x_0^{\mathrm{F}} - x\| : x \in X_*\}$. Then, for every $k \geq 0$, the following statements hold:*

a)

$$\min_{x\in\mathbb{R}^n}\left\{B_k\Theta_k(x) + \frac{1}{2}\|x - x_0^{\mathrm{F}}\|^2\right\} \geq B_k\Phi(y_k^{\mathrm{F}}) + \sum_{i=1}^{k}\frac{B_i(1-\sigma_{\mathrm{F}})}{2\lambda_{\mathrm{F}}}\|\tilde{y}_i^{\mathrm{F}} - \tilde{x}_{i-1}^{\mathrm{F}}\|^2; \quad (116)$$

b)

$$\|x_k^{\mathrm{F}} - x_*\| \leq R_0^{\mathrm{F}} + \sqrt{2\delta_0^{\mathrm{F}}C_{\mathrm{F}}}, \quad (117)$$

*where $C_{\mathrm{F}}$ is as in Theorem 4.5.*

**Proof**: a) The statement follows by induction. Relation (116) trivially holds for $k = 0$ since $B_0 = 0$. Then we assume the claim holds for some $k \geq 0$. For convenience, we denote

$$\beta_k = \sum_{i=1}^{k}\frac{B_i(1-\sigma)}{2\lambda}\|\tilde{y}_i - \tilde{x}_{i-1}\|^2. \quad (118)$$

It follows from Lemma C.3(a) and (b) that for every $u \in \mathbb{R}^n$ that

$$\begin{aligned}
B_k\Theta_k(u) + \frac{1}{2}\|u - x_0\|^2 &\geq B_k\Theta_k(x_k) + \frac{1}{2}\|x_k - x_0\|^2 + \frac{1+\mu B_k}{2}\|x_k - u\|^2 \\
&\geq B_k\Phi(y_k) + \beta_k + \frac{1+\mu B_k}{2}\|x_k - u\|^2, \quad (119)
\end{aligned}$$

where the second inequality follows from the inductive hypothesis. Using (112) and (119), we have

$$\begin{aligned}
B_{k+1}\Theta_{k+1}(u) + \frac{1}{2}\|u - x_0\|^2 - B_{k+1}\delta_k - b_k\theta_{k+1}(u) &\overset{(112)}{=} B_k\Theta_k(u) + \frac{1}{2}\|u - x_0\|^2 \\
&\overset{(119)}{\geq} B_k\Phi(y_k) + \beta_k + \frac{1+\mu B_k}{2}\|x_k - u\|^2 \geq B_k\theta_{k+1}(y_k) + \beta_k + \frac{\tau_k}{2}\|x_k - u\|^2,
\end{aligned}$$

where the last inequality follows from Lemmas C.3(c) and C.2(b). For $u \in \mathbb{R}^n$, define $\tilde{u} = B_{k+1}^{-1}(b_k u + B_k y_k)$. Rearranging the terms, we have

$$\begin{aligned} B_{k+1}\Theta_{k+1}(u) + \frac{1}{2}\|u - x_0\|^2 &\geq B_{k+1}\delta_k + b_k\theta_{k+1}(u) + B_k\theta_{k+1}(y_k) + \beta_k + \frac{\tau_k}{2}\|x_k - u\|^2 \\ &\overset{(62)}{\geq} B_{k+1}\theta_{k+1}(\tilde{u}) + \beta_k + B_{k+1}\delta_k + \frac{\tau_k B_{k+1}^2}{2b_k^2}\|\tilde{x}_k - \tilde{u}\|^2 \\ &= B_{k+1}\theta_{k+1}(\tilde{u}) + \beta_k + B_{k+1}\delta_k + \frac{B_{k+1}}{2\lambda}\|\tilde{x}_k - \tilde{u}\|^2. \end{aligned} \tag{120}$$

where the second inequality is due to the convexity of $\theta_{k+1}$ and (62) and the identity is due to Lemma C.2(a). It follows from the definition of $\theta_{k+1}$ in (111) that

$$\theta_{k+1}(u) \geq \theta_{k+1}(u) - \frac{\mu}{2}\|x - z_{k+1}\|^2 \overset{(111)}{=} \Gamma_k(z_{k+1}) - \frac{1}{2\lambda}\|z_{k+1} - \tilde{x}_k\|^2 + \langle u_{k+1}, u - z_{k+1}\rangle.$$

This inequality and (120) imply that

$$\begin{aligned} &B_{k+1}\Theta_{k+1}(u) + \frac{1}{2}\|u - x_0\|^2 \\ &\overset{(120)}{\geq} B_{k+1}\left(\Gamma_k(z_{k+1}) - \frac{1}{2\lambda}\|z_{k+1} - \tilde{x}_k\|^2 + \langle u_{k+1}, \tilde{u} - z_{k+1}\rangle + \delta_k + \frac{1}{2\lambda}\|\tilde{u} - \tilde{x}_k\|^2\right) + \beta_k. \end{aligned}$$

Minimizing over both sides of the above inequality, we obtain

$$\begin{aligned} &\min_{x\in\mathbb{R}^n}\left\{B_{k+1}\Theta_{k+1}(x) + \frac{1}{2}\|x - x_0\|^2\right\} \\ &\geq B_{k+1}\min_{x\in\mathbb{R}^n}\left\{\Gamma_k(z_{k+1}) - \frac{1}{2\lambda}\|z_{k+1} - \tilde{x}_k\|^2 + \langle u_{k+1}, x - z_{k+1}\rangle + \delta_k + \frac{1}{2\lambda}\|x - \tilde{x}_k\|^2\right\} + \beta_k \\ &\overset{(113)}{\geq} B_{k+1}\Phi(\tilde{y}_{k+1}) + \frac{B_{k+1}(1-\sigma)}{2\lambda}\|\tilde{y}_{k+1} - \tilde{x}_k\|^2 + \beta_k, \end{aligned}$$

where the last inequality follows by Lemma C.4. Finally, the target inequality (116) directly follows from the fact that $\Phi(\tilde{y}_{k+1}) \geq \Phi(y_{k+1})$ (see Step **3** of Algorithm 6) and the observation that $\beta_{k+1} = \beta_k + B_{k+1}(1-\sigma)\|\tilde{y}_{k+1} - \tilde{x}_k\|^2/(2\lambda)$ in view of (118).

b) It follows from (119) with $u = x_*$ that

$$B_k(\Theta_k(x_*) - \Phi(y_k)) \overset{(119)}{\geq} \frac{1+\mu B_k}{2}\|x_* - x_k\|^2 - \frac{1}{2}\|x_* - x_0\|^2 + \beta_k \geq \frac{1}{2}\|x_* - x_k\|^2 - \frac{1}{2}\|x_* - x_0\|^2.$$

Using the second inequality in Lemma C.3(c) with $x = x_*$ and Lemma C.2(e), we have

$$B_k(\Theta_k(x_*) - \Phi(y_k)) \leq B_k(\Phi(x_*) - \Phi(y_k) + \Delta_{k-1}) \leq B_k\Delta_{k-1} \overset{(110)}{\leq} \delta_0\sum_{i=0}^{k-1} B_{i+1}\alpha^i.$$

In view of the definition of $C_{\mathrm{F}}$ in Lemma C.2(c), the statement directly follows from combining the above two inequalities. ∎

We are now ready to prove Theorem 4.5, which directly follows from Lemmas C.2 and C.5.

**Proof of Theorem 4.5:** It follows from Lemma C.5(a),

$$B_{k+1}\Phi(y_{k+1}) + \sum_{i=1}^{k+1} \frac{B_i(1-\sigma)}{2\lambda}\|\tilde{y}_i - \tilde{x}_{i-1}\|^2 \overset{(116)}{\leq} B_{k+1}\Theta_{k+1}(x_*) + \frac{1}{2}\|x_* - x_0\|^2$$
$$\leq B_{k+1}\Phi(x_*) + \frac{1}{2}\|x_* - x_0\|^2 + B_{k+1}\Delta_k$$

where the second inequality is due to Lemma C.3(c) with $x = x_*$. Using Lemma C.2(e) yields

$$B_{k+1}(\Phi(y_{k+1}) - \Phi(x_*)) + \sum_{i=1}^{k+1} \frac{B_i(1-\sigma)}{2\lambda}\|\tilde{y}_i - \tilde{x}_{i-1}\|^2 \leq \frac{1}{2}\|x_* - x_0\|^2 + \delta_0 C_{\mathrm{F}}.$$

Therefore, (67)-(69) immediately follow. ∎

In the course of our analysis in Subsection 5.3 (see Proposition 5.10), we found it necessary to uniformly bound the distance from the prox center $\tilde{x}_k^{\mathrm{F}}$ to the minimum distance optimizer $x_*$. The following lemma provides a uniform upper bound on $\|\tilde{x}_k^{\mathrm{F}} - x_*\|$ over the iterations $k \geq 0$. Part 1 in the proof of Lemma C.6 below largely follows [32, Theorem 3.10].

**Lemma C.6.** *Suppose we choose $y_k^{\mathrm{F}} = \tilde{y}_k^{\mathrm{F}}$ in Step 3 of Algorithm 6 and $\sigma_{\mathrm{F}} < 1$, and suppose that $x_0^{\mathrm{F}} = 0$. Define $\bar{R}_0^{\mathrm{F}} = \max\{1, R_0^{\mathrm{F}}\}$, where $R_0^{\mathrm{F}} := \|x_0^{\mathrm{F}} - x_*\| = \min\{\|x_0^{\mathrm{F}} - x\| : x \in X_*\}$, where $X_*$ is the optimal solution set to* (49)*. Then, for every $k \geq 0$, we have*

$$\|\tilde{x}_k^{\mathrm{F}} - x_*\| \leq \mathcal{R}_{\mathrm{F}}, \tag{121}$$

*where $\mathcal{R}_{\mathrm{F}}$ is defined as*

$$\mathcal{R}_{\mathrm{F}} := \bar{R}_0^{\mathrm{F}}(1 + \sqrt{2\delta_0^{\mathrm{F}} C_{\mathrm{F}}})\left(\frac{2}{\sqrt{1-\sigma_{\mathrm{F}}}} + 1\right), \tag{122}$$

*and $C_{\mathrm{F}}$ is as in Theorem 4.5.*

**Proof**: We prove the claim in three parts. First, we provide an upper bound on $\|y_{k+1} - x_*\|$. Second, we combine the previous bound with Lemma C.5(b) to bound $\|\tilde{x}_k - x_*\|$.

**Part 1)** First, we show by induction that for every $k \geq 0$,

$$\|y_{k+1} - x_*\| \leq \frac{1}{B_{k+1}} \sum_{i=0}^{k} B_{i+1}\|\tilde{x}_i - y_{i+1}\| + R_0 + \sqrt{2\delta_0 C_{\mathrm{F}}}. \tag{123}$$

For $k = 0$, observe that $b_0 = B_1$ and $\tilde{x}_0 = x_0$ in view of (61) and (62), the claim (123) follows directly from the triangle inequality

$$\|y_1 - x_*\| \leq \|x_0 - y_1\| + \|x_0 - x_*\| = \frac{1}{B_1} \sum_{i=0}^{0} B_{i+1}\|\tilde{x}_i - y_{i+1}\| + R_0,$$

which proves the base case. We now perform the inductive step. First, applying the triangle inequality twice and using (62), we have

$$\|y_{k+1} - x_*\| \leq \|y_{k+1} - \tilde{x}_k\| + \|x_* - \tilde{x}_k\| \overset{(62)}{\leq} \|y_{k+1} - \tilde{x}_k\| + \frac{B_k}{B_{k+1}}\|x_* - y_k\| + \frac{b_k}{B_{k+1}}\|x_* - x_k\|.$$

It thus follows from Lemma C.5(b) that

$$\|y_{k+1} - x_*\| \overset{(117)}{\leq} \|y_{k+1} - \tilde{x}_k\| + \frac{B_k}{B_{k+1}}\|x_* - y_k\| + \frac{b_k}{B_{k+1}}(R_0 + \sqrt{2\delta_0 C_{\mathrm{F}}}).$$

Applying the inductive hypothesis (123) with $k+1$ replaced by $k$, we obtain

$$\|y_{k+1}-x_*\| \overset{(123)}{\leq} \|y_{k+1}-\tilde{x}_k\| + \frac{1}{B_{k+1}}\sum_{i=0}^{k-1} B_{i+1}\|\tilde{x}_i - y_{i+1}\| + \frac{B_k}{B_{k+1}}(R_0+\sqrt{2\delta_0 C_{\mathrm{F}}})$$
$$+\frac{b_k}{B_{k+1}}(R_0+\sqrt{2\delta_0 C_{\mathrm{F}}}) = \frac{1}{B_{k+1}}\sum_{i=0}^{k} B_{i+1}\|\tilde{x}_i - y_{i+1}\| + R_0 + \sqrt{2\delta_0 C_{\mathrm{F}}}.$$

Hence, we prove (123) holds for every $k \geq 0$.

It follows from (68) from Theorem 4.5 and (123) that

$$\|y_{k+1}-x_*\| \leq \left(\frac{\sqrt{\lambda}}{\sqrt{1-\sigma}B_{k+1}}\sum_{i=0}^{k}\sqrt{B_{i+1}}+1\right)(R_0+\sqrt{2\delta_0 C_{\mathrm{F}}}).$$

Since $\{B_k\}$ is increasing, we have

$$\|y_{k+1}-x_*\| \leq \left(\frac{\sqrt{\lambda}(k+1)}{\sqrt{1-\sigma}\sqrt{B_{k+1}}}+1\right)(R_0+\sqrt{2\delta_0 C_{\mathrm{F}}}) \leq \left(\frac{2}{\sqrt{1-\sigma}}+1\right)(R_0+\sqrt{2\delta_0 C_{\mathrm{F}}}), \quad (124)$$

where the second inequality follows from the fact that $B_{k+1} \geq \lambda(k+1)^2/4$ from Lemma C.2(c). Using the definition $\bar{R}_0 = \max\{1, R_0\}$, we note that $R_0 + \sqrt{2\delta_0 C_{\mathrm{F}}} \leq \bar{R}_0(1+\sqrt{2\delta_0 C_{\mathrm{F}}})$. Therefore, we conclude from (124) and the definition of $\mathcal{R}_{\mathrm{F}}$ in (122) that

$$\|y_{k+1}-x_*\| \leq \mathcal{R}_{\mathrm{F}}. \quad (125)$$

**Part 2)** Next, combining the triangle inequality with the fact that $(\alpha a+\beta b)^2 \leq (\alpha+\beta)(\alpha a^2+\beta b^2)$ for $a, b, \alpha, \beta \in \mathbb{R}_{++}$, we have

$$\|\tilde{x}_k - x_*\|^2 \overset{(62)}{\leq} \left(\frac{B_k}{B_{k+1}}\|y_k-x_*\| + \frac{b_k}{B_{k+1}}\|x_k-x_*\|\right)^2$$
$$\leq \left(\frac{B_k}{B_{k+1}}+\frac{b_k}{B_{k+1}}\right)\left(\frac{B_k}{B_{k+1}}\|y_k-x_*\|^2 + \frac{b_k}{B_{k+1}}\|x_k-x_*\|^2\right)$$
$$\overset{(125)}{\leq} \frac{B_k}{B_{k+1}}\mathcal{R}_{\mathrm{F}}^2 + \frac{b_k}{B_{k+1}}\|x_k-x_*\|^2 \overset{(117)}{\leq} \frac{B_k}{B_{k+1}}\mathcal{R}_{\mathrm{F}}^2 + \frac{b_k}{B_{k+1}}(R_0+\sqrt{2\delta_0 C_{\mathrm{F}}})^2 \leq \mathcal{R}_{\mathrm{F}}^2,$$

where the third inequality follows by Part 1 and the fact that $B_{k+1} = B_k + b_k$ (see (61)), the fourth inequality by Lemma C.5(b), and the final one by $R_0 + \sqrt{2\delta_0 C_{\mathrm{F}}} \leq \bar{R}_0(1+\sqrt{2\delta_0 C_{\mathrm{F}}}) \leq \mathcal{R}_{\mathrm{F}}$. ∎

# D Deferred Proofs for ACG and Restarted ACG

## D.1 FLOrA Analysis of Algorithm 1

In this subsection, we provide a self-contained analysis of Algorithm 1 by showing that it is an instance of the FLOrA framework.

Clearly the scalar sequences in Algorithm 6 and Algorithm 1 are equivalent with $b_k = a_j$, $B_k = A_j$, $\tau_k = \tau_j$, $\mu_{\mathrm{F}} = \mu$, and $\lambda_{\mathrm{F}} = 1/(2L)$. We can then restate Lemma C.2 for Algorithm 1.

**Lemma D.1.** *The following statements hold for every $j \geq 0$:*

*a)* $2La_j^2 = A_{j+1}\tau_j$;

*b)* $\tau_j = 1 + \mu A_j$;

*c)*

$$A_{j+1} \geq \frac{1}{2L}\max\left\{\frac{(j+1)^2}{4}, \left(1+\frac{1}{2}\sqrt{\frac{\mu}{2L}}\right)^{2j}\right\}.$$

We start by defining our lower model. On each iteration $j$, define

$$\Gamma_j(x) = \ell_g(x;\tilde{x}_j) + h(x) + \frac{2L+\mu}{2}\|u-\tilde{x}_j\|^2, \tag{126}$$

where $\Gamma_j$ is in fact the objective function in (17). To match the algorithm statements, we denote FLOrA iterates with $k$ and ACG iterates with $j$. Recalling the objective function $\psi(\cdot)$ defined in (14), we will show next that Algorithm 1 is an instance of the FLOrA framework (i.e., Algorithm 6) with the correspondence

$$\begin{gathered}\Phi(\cdot) = \psi(\cdot), \quad \Gamma_k^{\mathrm{F}}(\cdot) = \Gamma_j(\cdot), \quad \mathcal{A}_k^{\mathrm{F}} = \infty, \quad \delta_k^{\mathrm{F}} = \alpha_{\mathrm{F}} = 0, \quad \mu_{\mathrm{F}} = \mu, \quad \sigma_{\mathrm{F}} = 1/2; \\ \lambda_{\mathrm{F}} = \frac{1}{2L}, \quad y_k^{\mathrm{F}} = y_j, \quad z_k^{\mathrm{F}} = \tilde{y}_k^{\mathrm{F}} = \tilde{y}_j, \quad x_k^{\mathrm{F}} = x_j, \quad u_k^{\mathrm{F}} = u_j := 2L(\tilde{x}_{j-1} - \tilde{y}_j), \quad \hat{u}_k^{\mathrm{F}} = 0.\end{gathered} \tag{127}$$

First, We show that the lower model $\Gamma_j$ and $\tilde{y}_{j+1}$ satisfy (63) and (65) with $\mathcal{A}_k^{\mathrm{F}} = \infty$.

**Lemma D.2.** *Consider* $\Gamma_j(x)$ *defined in* (126)*. Then, the following statements hold:*

a) $\Gamma_j(u) \leq \psi(u) + L\|u-\tilde{x}_j\|^2$ *for every* $u \in \mathbb{R}^n$;

b) $\tilde{y}_{j+1} = \operatorname{argmin}_{u\in\mathbb{R}^n} \Gamma_j(u)$.

*Moreover, the two statements satisfy* (63) *and* (65) *with the correspondence* (127)*.*

**Proof**: a) The claim follows by the $\mu$-strong convexity of $g$ and the definition of $\Gamma_j$ in (126).

b) The claim follows directly from the definitions of $\tilde{y}_{j+1}$ in (17) and $\Gamma_j$ in (126).

Finally, it is easy to verify the final claim with the correspondence (127). ∎

Combining these properties with the $(L+\mu)$-smoothness and $\mu$-strong convexity of $g$, we show that $\Gamma_j$ satisfies the key inequality (64) in FLOrA.

**Lemma D.3.** *Consider* $\Gamma_j(x)$ *defined in* (126)*. Then, for every* $j \geq 0$*,*

$$\frac{1}{L}\left[\psi(\tilde{y}_{j+1}) + L\|\tilde{y}_{j+1}-\tilde{x}_j\|^2 - \Gamma_j(\tilde{y}_{j+1})\right] \leq \frac{1}{2}\|\tilde{y}_{j+1}-\tilde{x}_j\|^2. \tag{128}$$

*Moreover,* (128) *satisfies* (64) *with the correspondence* (127)*.*

**Proof**: Using the definition of $\Gamma_j$ in (126) and the $(L+\mu)$-smoothness of $g$, we have

$$\begin{aligned}\psi(\tilde{y}_{j+1}) + L\|\tilde{y}_{j+1}-\tilde{x}_j\| - \Gamma_j(\tilde{y}_{j+1}) &\overset{(126)}{=} g(\tilde{y}_{j+1}) - \ell_g(\tilde{y}_{j+1};\tilde{x}_j) - \frac{\mu}{2}\|\tilde{y}_{j+1}-\tilde{x}_j\|^2 \\ &\leq \frac{L+\mu}{2}\|\tilde{y}_{j+1}-\tilde{x}_j\|^2 - \frac{\mu}{2}\|\tilde{y}_{j+1}-\tilde{x}_j\|^2 = \frac{L}{2}\|\tilde{y}_{j+1}-\tilde{x}_j\|^2.\end{aligned}$$

Hence, (128) immediately follows. Finally, it is easy to verify the final claim with the correspondence (127). ∎

Finally, we show that the auxiliary sequence $\{x_{j+1}\}$ with update (19) is equivalent to the FLOrA sequence $\{x_{k+1}^{\mathrm{F}}\}$ with update (66).

**Lemma D.4.** *Choosing $u_{j+1}$ as in* (127)*, we can rewrite $x_{j+1}$ in* (19) *as*

$$x_{j+1} = \frac{1}{\tau_{j+1}} \left( \tau_j x_j - a_j u_{j+1} + \mu a_j \tilde{y}_{j+1} \right). \tag{129}$$

*Moreover,* (19) *is equivalent to* (66) *with the correspondence* (127)*.*

**Proof**: Using the definition of $x_{j+1}$ in (19) and Lemma D.1(a) and (b), we have

$$\begin{aligned} x_{j+1} &\overset{(19)}{=} \frac{1}{\tau_{j+1}} \left( 2La_j \tilde{y}_{j+1} + \tau_j x_j - \frac{2La_j^2}{A_{j+1}} x_j + \mu a_j \tilde{y}_{j+1} - 2L \frac{A_j a_j}{A_{j+1}} y_j \right) \\ &\overset{(16)}{=} \frac{1}{\tau_{j+1}} \left( \tau_j x_j + 2La_j \tilde{y}_{j+1} - 2La_j \tilde{x}_j + \mu a_j \tilde{y}_{j+1} \right) \\ &= \frac{1}{\tau_{j+1}} \left( \tau_j x_j - a_j u_{j+1} + \mu a_j \tilde{y}_{j+1} \right). \end{aligned}$$

where the second identity is due to (16) and the last one holds by our choice of $u_{j+1}$ in (127). Finally, it is easy to verify the last claim in view of (127) and (129). ∎

Having shown that Algorithm 1 is an instance of Algorithm 6, then Theorem 4.5 holds in the context of this subsection using the translation in (127). Therefore, we can present a simple proof of Lemma 2.1 based on the correspondence in (127).

**Proof of Lemma 2.1:** From the correspondence (127) and Theorem 4.5 with $R_0^{\mathrm{F}} = R_0$ and $B_k = A_j$, we obtain

$$\psi(y_j) - \psi_* \overset{(127)}{=} \Phi(y_k^{\mathrm{F}}) - \Phi_* \overset{(67),(127)}{\leq} \frac{R_0^2}{2A_j},$$

which proves (20).

The second claim follows by first noting that, by the definition of the gradient map,

$$\mathcal{G}_\psi^{(2L+\mu)^{-1}}(\tilde{x}_{j-1}) \overset{(13)}{=} (2L+\mu)(\tilde{x}_{j-1} - y_j). \tag{130}$$

Then, once again applying Theorem 4.5 with $R_0^{\mathrm{F}} = R_0$ and $B_k = A_j$ under the correspondence (127), we obtain

$$\|\mathcal{G}_\psi^{(2L+\mu)^{-1}}(\tilde{x}_{j-1})\| \overset{(130)}{=} (2L+\mu)\|y_j - \tilde{x}_{j-1}\| \overset{(127)}{=} (2L+\mu)\|y_k^{\mathrm{F}} - \tilde{x}_{k-1}^{\mathrm{F}}\| \overset{(68)(127)}{\leq} \frac{(2L+\mu)R_0}{\sqrt{LA_j}},$$

which proves (21). ∎

Similarly, since Algorithm 1 is an instance of FLOrA, the following lemma is a direct consequence of Lemmas C.3 and C.5(a) under the correspondence (127). The proof is omitted, since all results directly follow by substitution from (127).

**Lemma D.5.** *For all $j \geq 0$, let $\theta_j$ and $\Theta_j$ be as defined in* (22) *and* (23)*, respectively. Then, the following statements hold for every $j \geq 0$:*

a) *$\theta_{j+1}$ and $\Theta_{j+1}$ are $\mu$-strongly convex quadratic functions;*

b) *$x_j = \operatorname{argmin}_{x \in \mathbb{R}^n} \{A_j \Theta_j(x) + \|x - x_0\|^2/2\}$;*

c) *for all $x \in \operatorname{dom} \psi$, $\theta_{j+1}(x) \leq \psi(x)$ and $\Theta_{j+1}(x) \leq \psi(x)$;*

d)

$$A_j\psi(y_j) \leq \min_{u\in\mathbb{R}^n}\left\{A_j\Theta_j(u) + \frac{1}{2}\|u-x_0\|^2\right\}. \tag{131}$$

We are now ready to prove Lemma 2.2, which is the key result enabling the inner complexity bound in Proposition 5.3.

**Proof of Lemma 2.2:** a) It follows from Lemma D.5(d) that

$$\psi(y_j) \overset{(131)}{\leq} \min_{u\in\mathbb{R}^n}\left\{\Theta_j(u) + \frac{1}{2A_j}\|u-x_0\|^2\right\} \leq \Theta_j(\hat{x}_j) + \frac{1}{2A_j}\|\hat{x}_j - x_0\|^2. \tag{132}$$

Using Lemma D.5(a) and (24), we have for every $u \in \mathbb{R}^n$,

$$\psi(y_j) - \frac{1}{2A_j}\|\hat{x}_j - x_0\|^2 \leq \Theta_j(\hat{x}_j) \overset{(24)}{\leq} \Theta_j(u) - \frac{\mu}{2}\|u-\hat{x}_j\|^2.$$

Taking $u = y_j$ in the above inequality and using Lemma D.5(c), we obtain

$$\|y_j - \hat{x}_j\|^2 \leq \frac{1}{\mu A_j}\|\hat{x}_j - x_0\|^2.$$

Using the above inequality, the triangle inequality, and the fact that $(a+b)^2 \leq 2(a^2+b^2)$, we have

$$\|\hat{x}_j - x_0\|^2 \leq 2(\|\hat{x}_j - y_j\|^2 + \|y_j - x_0\|^2) \leq \frac{2}{\mu A_j}\|\hat{x}_j - x_0\|^2 + 2\|y_j - x_0\|^2.$$

Hence, (25) follows from the assumption that $A_j \geq 3/\mu$ and (132).

b) It follows from Lemma D.5(d) that for any $u \in \mathbb{R}^n$

$$\begin{aligned}\psi(y_j) + \frac{1}{2}\left(\mu + \frac{1}{A_j}\right)\|u-x_j\|^2 &\overset{(131)}{\leq} \min_{u\in\mathbb{R}^n}\left\{\Theta_j(u) + \frac{1}{2A_j}\|u-x_0\|^2\right\} + \frac{1}{2}\left(\mu + \frac{1}{A_j}\right)\|u-x_j\|^2 \\ &\leq \Theta_j(u) + \frac{1}{2A_j}\|u-x_0\|^2,\end{aligned}$$

where the second inequality follows from Lemma D.5(a) and (b). Taking $u = y_j$ in the above inequality and Lemma D.5(c), we have

$$\frac{1}{2}\left(\mu + \frac{1}{A_j}\right)\|y_j - x_j\|^2 \leq \Theta_j(y_j) - \psi(y_j) + \frac{1}{2A_j}\|y_j - x_0\|^2 \leq \frac{1}{2A_j}\|y_j - x_0\|^2,$$

and hence

$$\|y_j - x_j\|^2 \leq \frac{1}{1+A_j\mu}\|y_j - x_0\|^2 \leq \frac{1}{4}\|y_j - x_0\|^2$$

where the second inequality is due to $A_j \geq 3/\mu$. Finally, (26) immediately follows from the above inequality, the triangle inequality and the definition of $s_j$ in (24). ∎

## D.2 Deferred Proofs from Subsection 5.1

We begin by proving that Algorithm 2 is an instance of the FLOrA framework. The proof directly follows from substituting terms using the correspondence (70).

**Proof of Lemma 5.1:** Clearly, the scalar sequences $b_k$, $B_k$, and $\tau_k$ are equivalent. Similarly, $y_k^{\mathrm{F}} = w_k$ given the choice $\tilde{y}_{k+1}^{\mathrm{F}} = y_j$, and $\tilde{x}_k^{\mathrm{F}}$ in (62) is equivalent to $\tilde{v}_k$ in (27) given $y_k^{\mathrm{F}} = w_k$ and the choice of $x_k^{\mathrm{F}} = v_k$ in (70). Then we need to show the FLOrA conditions (63), (64), (65) hold and the update (66) is equivalent to (30).

First, note that $\Theta_j$ is $(\mu_f + \lambda^{-1})$-strongly convex in view of Lemma D.5(a) and the choice of $\mu_{\mathrm{F}}$ in (28), matching the condition on $\Gamma_k^{\mathrm{F}}$ in Step **2** of Algorithm 6. Condition (63) holds by Lemma D.5(c) and the definition of $\psi$ in (28).

Second, the first relation in (65) holds by Lemma D.5(b) in view of $\tilde{x}_k^{\mathrm{F}} = \tilde{v}_k = x_0$ (see (28) and (70)) and the choices $z_{k+1}^{\mathrm{F}} = x_j$, $\Gamma_k^{\mathrm{F}} = \Theta_j$, and $\mathcal{A}_k^{\mathrm{F}} = A_j$ given in (70). Similarly, the second relation in (65) follows from the definition of $s_j$ in (24).

Third, we prove the equivalence of (64) and (29). By the correspondence (70) and the relation $\tilde{x}_k^{\mathrm{F}} = \tilde{v}_k = x_0$ noted above, we have

$$\begin{aligned}
&\|\lambda_{\mathrm{F}} \hat{u}_{k+1}^{\mathrm{F}}\|^2 + 2\lambda_{\mathrm{F}} \left[\Phi(\tilde{y}_{k+1}^{\mathrm{F}}) + \frac{1}{2\lambda_{\mathrm{F}}}\|\tilde{x}_k^{\mathrm{F}} - \tilde{y}_{k+1}^{\mathrm{F}}\|^2 - \Gamma_k^{\mathrm{F}}(z_{k+1}^{\mathrm{F}})\right] \\
&\overset{(70)}{=} \|\lambda s_j\|^2 + 2\lambda \left[\phi(y_j) + \frac{1}{2\lambda}\|x_0 - y_j\|^2 - \Theta_j(x_j)\right] \overset{(29)}{\leq} \sigma\|y_j - x_0\|^2 \overset{(70)}{=} \sigma_{\mathrm{F}}\|\tilde{y}_{k+1}^{\mathrm{F}} - \tilde{x}_k^{\mathrm{F}}\|^2
\end{aligned}$$

which proves (64) is equivalent to (29).

Finally we verify the equivalence of (66) and (30). Given our choice of $z_{k+1}^{\mathrm{F}} = x_j$ and the relation $\tilde{x}_k^{\mathrm{F}} = \tilde{v}_k = x_0$, we observe

$$u_{k+1}^{\mathrm{F}} \overset{(66)}{=} \hat{u}_{k+1}^{\mathrm{F}} + \frac{\tilde{x}_k^{\mathrm{F}} - z_{k+1}^{\mathrm{F}}}{\lambda} \overset{(70)}{=} s_j + \frac{x_0 - x_j}{\lambda} \overset{(24)}{=} s_j + \frac{A_j}{\lambda} s_j = \frac{\lambda + A_j}{\lambda} s_j,$$

which validates our choice of $u_{k+1}^{\mathrm{F}}$ in (70). Then with the choices $x_k^{\mathrm{F}} = v_k$, $z_{k+1}^{\mathrm{F}} = x_j$, $\mu_{\mathrm{F}} = \mu_f$, and $u_{k+1}^{\mathrm{F}} = \lambda^{-1}(A_j + \lambda)s_j$ given in (70), we can rewrite (30) as

$$v_{k+1} \overset{(30),(70)}{=} \frac{1}{\tau_{k+1}} \left(\tau_k x_k^{\mathrm{F}} + b_k \mu_{\mathrm{F}} z_{k+1}^{\mathrm{F}} - b_k u_{k+1}^{\mathrm{F}}\right) \overset{(66)}{=} x_{k+1}^{\mathrm{F}}.$$

Therefore, Algorithm 2 is an instance of the FLOrA framework. ∎

Using the results from Subsection 2.1 and Appendix D.1, we are now ready to prove Proposition 5.3, which connects the inner ACG subroutine with the outer termination condition.

**Proof of Proposition 5.3:** By Lemma D.1(c) with $L = L_f - \mu_f$ and $\mu = \lambda^{-1} + \mu_f$ (see (28)), (71) implies that $A_j \geq 5\lambda/\sigma$, with the condition on $\lambda$ ensuring that the log term is non-negative. Using Lemma 2.2(a) and (b) with $\mu = \mu_f + \lambda^{-1}$, we have

$$\begin{aligned}
&\|\lambda s_j\|^2 + 2\lambda[\psi(y_j) - \Theta_j(x_j)] \overset{(25),(26)}{\leq} \frac{9\lambda^2\|y_j - x_0\|^2}{4A_j^2} + \frac{2\lambda(\mu_f + \lambda^{-1})}{A_j(\mu_f + \lambda^{-1}) - 2}\|y_j - x_0\|^2 \\
&\leq \frac{9\lambda^2\|y_j - x_0\|^2}{4A_j^2} + \frac{2}{A_j\lambda^{-1} - 2}\|y_j - x_0\|^2 \leq \left(\frac{\sigma^2}{10} + \frac{2\sigma}{3}\right)\|y_j - x_0\|^2 \leq \sigma\|y_j - x_0\|^2,
\end{aligned}$$

where the third inequality follows from the facts that $A_j \geq 5\lambda/\sigma$ and $\sigma \in (0, 1)$. ∎

# E Deferred Proofs for I-ALM and I-FALM

## E.1 Deferred Proofs for I-ALM

**Proof of Proposition 5.4** Set $L$ and $\mu$ as in (33) define $\Phi(\cdot) = \mathcal{L}_\rho(\cdot, \lambda_k)$, $\gamma = \varepsilon_k/(4D^2)$, $\eta = (2L+\mu)^{-1}$, and $\bar{x} = x_k$. Using Lemma B.1(c) and noting $\phi_\gamma(\cdot) = \psi(\cdot)$ in light of (33), requiring $\|\mathcal{G}_\psi^{(2L+\mu)^{-1}}(\tilde{x}_k)\| \leq \varepsilon_k/4D$ guarantees that $\|\mathcal{G}_{\mathcal{L}_\rho(\cdot,\lambda_k)}^{(2L+\mu)^{-1}}(\tilde{x}_k)\| \leq \varepsilon_k/2D$ as in Step **1**. Applying Lemma 2.1 to $\psi$ as in (33) together with Lemma D.1(c), each call to Algorithm 1 in Step **1** takes

$$\tilde{\mathcal{O}}\left(1 + \frac{\sqrt{L}}{\sqrt{\mu}}\right) = \tilde{\mathcal{O}}\left(1 + \frac{D(\sqrt{L_f} + \sqrt{\rho}\|A\|)}{\sqrt{\varepsilon_k}}\right)$$

ACG iterations to guarantee $\|\mathcal{G}_\psi^{(2L+\mu)^{-1}}(\tilde{x}_k)\| \leq \varepsilon_k/4D$. The claim follows by taking the uniform lower bound $\varepsilon_k \geq \sigma\rho\varepsilon^2/2$. ∎

**Proof of Lemma 5.5** a) The claim immediately follows from the update (34) and the definition of $\mathcal{L}(x, \lambda)$ in (3).

b) The inequality directly follows by the definition of the dual $d(\nu)$ in (3), which implies $-d(\nu) \geq -\mathcal{L}(x, \nu)$ for all $x \in \mathbb{R}^n$. ∎

**Proof of Corollary 3.2** Noting that $\delta_k^{\mathrm{L}} = \varepsilon_0\alpha^k$ (see (75)) is summable with

$$\sum_{i=0}^{\infty} \delta_i^{\mathrm{L}} = \sum_{i=0}^{\infty} \varepsilon_0\alpha^i = \frac{\varepsilon_0}{1-\alpha}.$$

Hence, applying Lemma C.1 to $\Phi(\cdot) = -d(\cdot)$ with $R_0^{\mathrm{L}} = R_\Lambda$ and $C_\delta = \varepsilon_0/(1-\alpha)$, we have

$$\|\lambda_k - \lambda_*\| \overset{(75)}{=} \|x_k^{\mathrm{L}} - x_*\| \overset{(75),(108)}{\leq} R_\Lambda + \sqrt{\frac{2\rho\varepsilon_0}{1-\alpha}} = R_\Lambda + \sqrt{\frac{2}{1-\alpha}} \leq R_\Lambda + D\sqrt{\frac{2}{1-\alpha}}, \tag{133}$$

where the second equality follows by the choice $\rho = \varepsilon^{-1} = \varepsilon_0^{-1}$ in Theorem 3.1 and the last inequality is due to Assumption 2(d). It follows from the triangle inequality and $\lambda_0 = 0$ that

$$\|\lambda_k\| \leq \|\lambda_k - \lambda_*\| + \|\lambda_*\| = \|\lambda_k - \lambda_*\| + R_\Lambda \overset{(133)}{\leq} 2R_\Lambda + \zeta D, \tag{134}$$

where $\zeta = \sqrt{2/(1-\alpha)}$. Suppose that $(x_k, \lambda_k)$ is an $\varepsilon/[2R_\Lambda + (1+\zeta)D]$-solution to (2), then by Lemma B.3, we have

$$|\phi(x_k) - \hat{\phi}_*| \overset{(102)}{\leq} \max\{R_\Lambda, \|\lambda_k\| + D\}\frac{\varepsilon}{2R\lambda + (1+\zeta)D} \overset{(134)}{\leq} \frac{\varepsilon}{2R_\Lambda + (1+\zeta)D}(2R_\Lambda + (1+\zeta)D) = \varepsilon,$$

where the second inequality follows from the fact that $R_\Lambda \leq 2R_\Lambda + \zeta D$. Then, by Theorem 3.1 with $\varepsilon$ replaced by $\varepsilon/(2R_\Lambda + (1+\zeta)D) \leq \varepsilon$, the iteration-complexity for $(x_k, \lambda_k)$ to guarantee $|\phi(x_k) - \hat{\phi}_*| \leq \varepsilon$ and $\|Ax_k - b\| \leq \varepsilon$ is given by (36). ∎

## E.2 Deferred Proofs for I-FALM

### E.2.1 Proof of Lemma 3.3

By subdifferential calculus, we can show

$$v \in \partial\tilde{\mathcal{L}}(\cdot, \lambda)(x) = \partial\left(\mathcal{L}(\cdot, \lambda) + \frac{\gamma_p}{2}\|\cdot - x_0\|^2\right)(x) = \partial\mathcal{L}(\cdot, \lambda)(x) + \gamma_p(x - x_0),$$

hence rearranging yields

$$v' := v - \gamma_p(x - x_0) \in \partial \mathcal{L}(\cdot, \lambda)(x).$$

Then, applying the triangle inequality, Assumption 2(d), and the choice $\gamma_p = \varepsilon/(2D)$ we have

$$\|v'\| \leq \|v\| + \gamma_p \|x - x_0\| \leq \frac{\varepsilon}{2} + \frac{\varepsilon D}{2D} \leq \varepsilon.$$

### E.2.2 Proof of Lemma 5.7

Let $\lambda_* \in \Lambda_*$ be the optimal multiplier achieving $R_\Lambda$. Since $-\tilde{d}$ is $\gamma_d$-strongly convex with minimizer $\tilde{\lambda}_*$, we have

$$\begin{aligned} -\tilde{d}(\lambda_*) + \tilde{d}(\tilde{\lambda}_*) &\geq \frac{\gamma_d}{2}\|\lambda_* - \tilde{\lambda}_*\|^2 = \frac{\gamma_d}{2}\|\lambda_* - \lambda_0\|^2 + \gamma_d\langle \lambda_* - \lambda_0, \lambda_0 - \tilde{\lambda}_*\rangle + \frac{\gamma_d}{2}\|\lambda_0 - \tilde{\lambda}_*\|^2 \\ &\geq \frac{\gamma_d}{2}R_\Lambda^2 - \gamma_d R_\Lambda R_{\tilde{\Lambda}} + \frac{\gamma_d}{2}R_{\tilde{\Lambda}}^2, \end{aligned} \tag{135}$$

where the second inequality follows from the Cauchy-Schwarz inequality and the facts that $R_\Lambda = \|\lambda_0 - \lambda_*\|$ and $R_{\tilde{\Lambda}} = \|\lambda_0 - \tilde{\lambda}_*\|$. Define $u(\lambda) = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \mathcal{L}(x, \lambda)$, then by the definitions of $d$ and $\tilde{d}$ in (3) and (38), respectively, we have

$$\tilde{d}(\lambda) \overset{(38)}{\leq} \mathcal{L}(u(\lambda), \lambda) + \frac{\gamma_p}{2}\|u(\lambda) - x_0\|^2 - \frac{\gamma_d}{2}\|\lambda - \lambda_0\|^2 \overset{(3)}{=} d(\lambda) + \frac{\gamma_p}{2}\|u(\lambda) - x_0\|^2 - \frac{\gamma_d}{2}\|\lambda - \lambda_0\|^2. \tag{136}$$

Applying (39) with $\lambda = \lambda_*$ and (136) with $\lambda = \tilde{\lambda}_*$, we obtain

$$-\tilde{d}(\lambda_*) + \tilde{d}(\tilde{\lambda}_*) \overset{(39),(136),}{\leq} -d(\lambda_*) + \frac{\gamma_d}{2}\|\lambda_* - \lambda_0\|^2 + d(\tilde{\lambda}_*) + \frac{\gamma_p}{2}\|u(\tilde{\lambda}_*) - x_0\|^2 - \frac{\gamma_d}{2}\|\tilde{\lambda}_* - \lambda_0\|^2.$$

Plugging $R_\Lambda = \|\lambda_0 - \lambda_*\|$ and $R_{\tilde{\Lambda}} = \|\lambda_0 - \tilde{\lambda}_*\|$, and using Assumption 2(d), we have

$$-\tilde{d}(\lambda_*) + \tilde{d}(\tilde{\lambda}_*) \leq -d(\lambda_*) + d(\tilde{\lambda}_*) + \frac{\gamma_d}{2}R_\Lambda^2 - \frac{\gamma_d}{2}R_{\tilde{\Lambda}}^2 + \frac{\gamma_p}{2}D^2 \leq \frac{\gamma_d}{2}R_\Lambda^2 - \frac{\gamma_d}{2}R_{\tilde{\Lambda}}^2 + \frac{\gamma_p}{2}D^2. \tag{137}$$

where the second inequality follows from $d(\tilde{\lambda}_*) \leq d(\lambda_*)$. Combining the lower bound (135) and upper bound (137) on $-\tilde{d}(\lambda_*) + \tilde{d}(\tilde{\lambda}_*)$, we obtain

$$\frac{\gamma_d}{2}R_\Lambda^2 - \frac{\gamma_d}{2}R_{\tilde{\Lambda}}^2 + \frac{\gamma_p}{2}D^2 \overset{(137)}{\geq} -\tilde{d}(\lambda_*) + \tilde{d}(\tilde{\lambda}_*) \overset{(135)}{\geq} \frac{\gamma_d}{2}R_\Lambda^2 - \gamma_d R_\Lambda R_{\tilde{\Lambda}} + \frac{\gamma_d}{2}R_{\tilde{\Lambda}}^2,$$

which, by rearranging, yields

$$R_{\tilde{\Lambda}}^2 - R_{\tilde{\Lambda}} R_\Lambda - \frac{\gamma_p}{2\gamma_d}D^2 \leq 0. \tag{138}$$

If $\gamma_p = \varepsilon/(2D)$ and $\gamma_d = C_0\varepsilon/(R_{\tilde{\Lambda}})$ for some constant $C_0 > 0$, then $\gamma_p/\gamma_d = R_{\tilde{\Lambda}}/(2C_0 D)$. Then, we can rewrite (138) as

$$R_{\tilde{\Lambda}}^2 - R_{\tilde{\Lambda}}\left(R_\Lambda + \frac{D}{4C_0}\right) \leq 0.$$

Clearly, $R_{\tilde{\Lambda}}$ attains its extremal values when the LHS equals 0. Taking the nonzero solution $R_\Lambda + D/(4C_0)$ yields the claim (82).

### E.2.3 Proof of Proposition 5.8

Set $L$ and $\mu$ as in (41) and define $\Phi(\cdot) = \mathcal{L}_\rho(\cdot, \tilde{\nu}_k)$, $\gamma = \varepsilon_k/(4D^2)$, $\eta = (2L+\mu)^{-1}$, and $\bar{x} = x_k$. Using Lemma B.1(c) and noting $\phi_\gamma(\cdot) = \psi(\cdot)$ in light of (41), requiring $\|\mathcal{G}_\psi^{(2L+\mu)^{-1}}(\tilde{x}_k)\| \leq \varepsilon_k/4D$ guarantees that $\|\mathcal{G}_{\mathcal{L}_\rho(\cdot,\lambda_k)}^{(2L+\mu)^{-1}}(\tilde{x}_k)\| \leq \varepsilon_k/2D$ as in Step **2**. Applying Lemma 2.1 to $\psi$ and using Lemma D.1(c) with $L$ and $\mu$ as in (41), we show that each call to Algorithm 1 takes

$$\tilde{\mathcal{O}}\left(1 + \frac{\sqrt{L}}{\sqrt{\mu}}\right) = \tilde{\mathcal{O}}\left(1 + \frac{\sqrt{L_f} + \sqrt{\rho}\|A\|}{\sqrt{\gamma_p + \varepsilon_k/(4D^2)}}\right)$$

ACG iterations to guarantee $\|\mathcal{G}_\psi^{(2L+\mu)^{-1}}(\tilde{x}_k)\| \leq \varepsilon_k/(4D)$. The claim follows from the trivial lower bound $\gamma_p + \varepsilon_k/(4D^2) \geq \gamma_p = \varepsilon/(2D)$.

### E.2.4 Proof of Lemma 5.9

a) We first note that

$$\nabla\left(-\tilde{\mathcal{L}}(x_{k+1}, \cdot) + \frac{1}{2\rho}\|\cdot - \tilde{\nu}_k\|^2\right)(\lambda_{k+1}) = \gamma_d\lambda_{k+1} - (Ax_{k+1} - b) + \frac{\lambda_{k+1} - \tilde{\nu}_k}{\rho} \overset{(42)}{=} \gamma_d\lambda_{k+1}.$$

It thus follows from the $(\rho^{-1} + \gamma_d)$-strong convexity of $-\tilde{\mathcal{L}}(x_{k+1}, \cdot) + \|\cdot - \tilde{\nu}_k\|^2/(2\rho)$ that for every $\nu \in \mathbb{R}^m$,

$$\begin{aligned}\Gamma_k^\lambda(\nu) &= -\tilde{\mathcal{L}}(x_{k+1}, \lambda_{k+1}) + \frac{1}{2\rho}\|\lambda_{k+1} - \tilde{\nu}_k\|^2 + \langle\gamma_d\lambda_{k+1}, \nu - \lambda_{k+1}\rangle + \frac{1+\gamma_d\rho}{2\rho}\|\nu - \lambda_{k+1}\|^2 \\ &\leq -\tilde{\mathcal{L}}(x_{k+1}, \nu) + \frac{1}{2\rho}\|\nu - \tilde{\nu}_k\|^2.\end{aligned}$$

Hence, this statement immediately follows from the definition of $\tilde{d}$ in (38).

b) Since $\Gamma_k^\lambda(\nu)$ is a quadratic function, it is easy to verify that $\hat{\lambda}_{k+1}$ as in (84) is the solution to $\min\{\Gamma_k^\lambda(\nu) : \nu \in \mathbb{R}^m\}$. Also, it is straightforward to verify (87) by computation.

c) The claim follows directly from (43), $u_{k+1} = \rho^{-1}(\tilde{\nu}_k - \hat{\lambda}_{k+1})$, and the definition of $\hat{\lambda}_{k+1}$ in (84).

d) The claim follows from the requirement $\alpha \leq (1+\sqrt{\rho\gamma_d})^{-2}$ in the initialization of Algorithm 4 and Lemma C.2(d) with $C_{\mathrm{F}} = C$ in view of the definition of $C$ in (44) and the correspondence $\alpha_{\mathrm{F}} = \alpha$, $\mu_{\mathrm{F}} = \gamma_d$, and $\lambda_{\mathrm{F}} = \rho$ in (86).

### E.2.5 Proof of Corollary 3.5

By the parameters chosen in Theorem 3.4, Lemma 5.9 and Proposition 5.10 imply that Algorithm 4 is an instance of the FLOrA framework under the correspondence (86). Then, applying Lemma C.6 with $\Phi(\cdot) = -\tilde{d}(\cdot)$ noting that $\mathcal{R}_{\mathrm{F}} = \mathcal{R}$ (where $\mathcal{R}$ is as in (44)), we obtain for any $k \geq 1$,

$$\|\tilde{\nu}_{k-1} - \tilde{\lambda}_*\| \overset{(86)}{=} \|\tilde{x}_{k-1}^{\mathrm{F}} - x_*\| \overset{(86),(121)}{\leq} \mathcal{R}. \tag{139}$$

Then, it follows from the triangle inequality that

$$\begin{aligned}\|\lambda_k\| &\leq \|\lambda_k - \tilde{\nu}_{k-1}\| + \|\tilde{\nu}_{k-1} - \tilde{\lambda}_*\| + \|\tilde{\lambda}_*\| \overset{(42)}{=} \rho\|Ax_k - b\| + \|\tilde{\nu}_{k-1} - \tilde{\lambda}_*\| + \|\tilde{\lambda}_*\| \\ &\overset{(139)}{\leq} \rho\|Ax_k - b\| + \mathcal{R} + \|\tilde{\lambda}_*\| \leq \rho\|Ax_k - b\| + 2\mathcal{R},\end{aligned} \tag{140}$$

where the last inequality follows from the definition of $\mathcal{R}$ in (44) and the choice $\lambda_0 = 0$.

Suppose that $(x_k, \lambda_k)$ is an $\varepsilon$-primal-dual solution to (2). Then, by Lemma B.3 the absolute primal gap is bounded from above by

$$|\phi(x_k) - \hat{\phi}_*| \overset{(102)}{\leq} \max\{(D + \|\lambda_k\|), R_\Lambda\}\varepsilon \overset{(140)}{\leq} (D + \rho\|Ax_k - b\| + 2\mathcal{R})\varepsilon \overset{(4)}{\leq} (D + 2\mathcal{R})\varepsilon + \rho\varepsilon^2 \leq 2(D + \mathcal{R})\varepsilon,$$

where the second inequality follows from $R_\Lambda \leq (D + \rho\|Ax_k - b\| + 2\mathcal{R})$ and the final inequality follows by the condition $\rho\varepsilon = 4\sigma\rho\varepsilon \leq 1$ and Assumption 2(d).

Therefore, (48) follows by substituting $\varepsilon = \varepsilon_g/(2(D + \mathcal{R}))$ into Theorem 3.4 and using the fact that $\mathcal{R} = \mathcal{O}(\hat{R}_\Lambda + D)$ under the parameter settings of Theorem 3.4 and noting that $\|Ax_k - b\| \leq \varepsilon_g \leq \varepsilon$.