# An Accelerated Inexact Dampened Augmented Lagrangian Method for Linearly-Constrained Nonconvex Composite Optimization Problems*

Weiwei Kong† and Renato D.C. Monteiro‡

October 21, 2021

**Abstract**

This paper proposes and analyzes an accelerated inexact dampened augmented Lagrangian (AIDAL) method for solving linearly-constrained nonconvex composite optimization problems. Each iteration of the AIDAL method consists of: (i) inexactly solving a dampened proximal augmented Lagrangian (AL) subproblem by calling an accelerated composite gradient (ACG) subroutine; (ii) applying a dampened and under-relaxed Lagrange multiplier update; and (iii) using a novel test to check whether the penalty parameter of the AL function should be increased. Under several mild assumptions involving the dampening factor and the under-relaxation constant, it is shown that the AIDAL method generates an approximate stationary point of the constrained problem in $\mathcal{O}(\varepsilon^{-5/2}\log\varepsilon^{-1})$ iterations of the ACG subroutine, for a given tolerance $\varepsilon > 0$. Numerical experiments are also given to show the computational efficiency of the proposed method.

## 1 Introduction

This paper presents an accelerated inexact dampened augmented Lagrangian (AIDAL) method for finding approximate stationary points of the linearly constrained nonconvex composite optimization (NCO) problem

$$\min_z \left\{\phi(u) := f(z) + h(z) : Az = b\right\}, \tag{1}$$

where $A$ is a linear operator, $h$ is a proper closed convex and Lipschitz continuous function with compact domain, and $f$ is a (possibly) nonconvex differentiable function on the domain of $h$ with a Lipschitz continuous gradient. More specifically, the AIDAL method is based on the $\theta$-dampened augmented Lagrangian (AL) function

$$\mathcal{L}_c^\theta(z;p) := \phi(z) + (1-\theta)\langle p, Az-b\rangle + \frac{c}{2}\|Az-b\|^2 \quad \forall c>0, \quad \forall \theta\in(0,1), \tag{2}$$

*The first author has been supported by (i) the US Department of Energy (DOE) and UT-Battelle, LLC, under contract DE-AC05-00OR22725, (ii) the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration, and (iii) the IDEaS-TRIAD Fellowship (NSF Grant CCF-1740776).

†Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37830. wwkong92@gmail.com

‡School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205. monteiro@isye.gatech.edu

and it performs the following updates to generate its $k^{\text{th}}$ iterate: given $(z_{k-1}, p_{k-1})$ and $(\lambda, c_k)$, compute

$$z_k \approx \underset{u}{\operatorname{argmin}} \left\{ \lambda \mathcal{L}^{\theta}_{c_k}(u; p_{k-1}) + \frac{1}{2}\|u - z_{k-1}\|^2 \right\}, \tag{3}$$

$$p_k = (1-\theta)p_{k-1} + \chi c_k (A z_k - b), \tag{4}$$

where $\chi$ is an under-relaxation parameter in $(0,1)$ and $z_k$ is a suitably chosen approximate solution of the composite problem underlying (3). In addition, the AIDAL method introduces a novel approach for updating the penalty parameter $c_k$ between iterations and uses an accelerated composite gradient (ACG) method applied to (3) obtain the aforementioned point $z_k$.

Under a suitable choice of $\lambda$ and the following Slater-like assumption:

$$\exists \bar{z} \in \operatorname{int}(\operatorname{dom} h) \text{ such that } A\bar{z} = b, \tag{5}$$

where $\operatorname{int}(\operatorname{dom} h)$ denotes the interior of the domain of $h$, it is shown that, for any tolerance pair $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$, the AIDAL method obtains a pair $([\hat{z}, \hat{p}], \hat{v})$ satisfying

$$\hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}) + A^*\hat{p}, \quad \|\hat{v}\| \le \hat{\rho}, \quad \|A\hat{z} - b\| \le \hat{\eta}. \tag{6}$$

in $\mathcal{O}((\hat{\eta}^{-5/2} + \hat{\eta}^{-1/2}\hat{\rho}^{-2}) \log \hat{\eta}^{-1})$ ACG iterations. Moreover, this iteration complexity is obtained without requiring that the initial point $z_0$ (in the domain of $h$) be feasible with respect to the linear constraint, i.e., $Az_0 = b$. Another contribution from this analysis is that the sequence of Lagrange multipliers is shown to be bounded by a constant that is independent of $\hat{\rho}$ and $\hat{\eta}$.

*Related Works.* To condense our discussion, we let $\varepsilon = \hat{\rho} = \hat{\eta}$ denote a common tolerance parameter and restrict our attention to works that establish iteration complexity bounds for obtaining approximate stationary points of (1). For an overview of papers that focus on asymptotic convergence of a proposed method, see the excellent discussion in [18, Section 2].

One popular class of methods for obtaining stationary points of (1) is the penalty method, which consists of solving a sequence of unconstrained subproblems containing an objective function that penalizes a violation of the constraints through a positively weighted penalty term. Papers [9,12] present an $\mathcal{O}(\varepsilon^{-3})$ iteration complexity of a quadratic penalty method without any regularity assumptions on the linear constraint. In a follow-up work, paper [10] presents an $\mathcal{O}(\varepsilon^{-3}\log \varepsilon^{-1})$ iteration complexity of a similar quadratic penalty method in which its parameters are chosen in an adaptive and numerically efficient manner. Paper [18] is the first to present a penalty-based method with an improved complexity of $\mathcal{O}(\varepsilon^{-5/2}\log \varepsilon^{-1})$ under the assumption that the domain of $h$ is compact and assumption (5) holds. Paper [16] extends this previous work by presenting a hybrid penalty/AL-based method that also obtains the same aforementioned complexity. Finally, papers [26] and [15] respectively present $\mathcal{O}(\varepsilon^{-3}\log \varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^{-5/2}\log \varepsilon^{-1})$ iteration complexities of some AL-based methods that perform under-relaxed Lagrange multiplier updates only when the penalty parameter is updated. It is worth noting that when these methods initialize the penalty parameter on the order of $\mathcal{O}(1)$, they perform a $\mathcal{O}(\log \varepsilon^{-1})$ number of multiplier updates of the form (4) with $\theta = 0$, $\chi = \chi_k$, and $\chi_k$ approaching 0. It is also worth noting that if the initial penalty parameter of these method is chosen sufficiently large, then they never perform multiplier updates, and hence behave more like penalty methods.

Another popular class of methods is the proximal AL (PAL) method, which primarily consists of the updates in (3) and (4). The analysis of AL/PAL-based methods for the case where $\phi$ is convex is already well-established (see, for example, [1, 2, 13, 14, 19, 20, 24, 25, 27]), so we make no more mention of it here. Instead, we review papers that present an iteration complexity of an AL/PAL-based method for the case where $\phi$ is nonconvex. Paper [6] presents an $\mathcal{O}(\varepsilon^{-4})$ iteration complexity of an unaccelerated PAL method under the strong assumption that the initial point $z_0$ is feasible, i.e., $Az_0 = b$, and where $\theta \in (0, 1]$ and $\chi = 1$. Paper [22] presents $\mathcal{O}(\varepsilon^{-3} \log \varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^{-5/2} \log \varepsilon^{-1})$ iteration complexities of an accelerated inexact PAL method for the general case and the case where (5) holds, respectively, and removes the requirement that the initial point be feasible. Finally, papers [11, 21] present an $\mathcal{O}(\varepsilon^{-3} \log \varepsilon^{-1})$ iteration complexity for the special case of $(\chi, \theta) = (1, 0)$, which corresponds to a full multiplier update under the classical AL function.

Aside from penalty and AL/PAL-based methods, we mention few others that are of interest. Paper [3] presents an $\mathcal{O}(\varepsilon^{-3})$ iteration complexity of a primal-dual proximal point scheme for generating a point *near* an approximate stationary point under some strong conditions on the initial point. Papers [28, 29] present an $\mathcal{O}(\varepsilon^{-2})$ iteration complexity of a primal-dual first-order algorithm for solving (1) when $h$ is the indicator function of a box (in [29]), or more generally, a polyhedron (in [28]). Paper [7] presents an $\mathcal{O}(\varepsilon^{-6})$ iteration complexity of a penalty-ADMM method that solves an equivalent reformulation of (1), under the assumption that the initial point $z_0$ is feasible, the tolerance $\varepsilon$ is sufficiently small, and $A$ has full row rank.

*Contributions.* We now emphasize how the proposed AIDAL method improves on other state-of-the-art AL-based works. First, in contrast to the $\mathcal{O}(\varepsilon^{-5/2} \log \varepsilon^{-1})$ complexity PAL method in [22] where $\lambda \to 0$ as $\theta \to 0$, it chooses $\lambda$ in update (3) independent of $\theta$. Moreover, as $\theta \to 0$ the AIDAL method becomes a penalty method with stepsize $\lambda$ (see Subsection 2.2) while the method in [22] becomes a PAL method with arbitrarily small stepsize. Second, it improves on the $\mathcal{O}(\varepsilon^{-5/2} \log \varepsilon^{-1})$ AL-based method in [15] in two significant ways: (i) it performs the multiplier update (4) after every inexact prox update as opposed to only when the penalty parameter is updated; and (ii) it chooses a constant under-relaxation parameter $\chi$ for the update (4) as opposed to in [15] which chooses a under-relaxation parameter that (linearly) tends to zero as the number of penalty parameter updates increases. Finally, it improves upon the $\mathcal{O}(\varepsilon^{-3} \log \varepsilon^{-1})$ classic PAL method in [21] by an $\mathcal{O}(\varepsilon^{-1/2})$ factor through only a *small* perturbation of the classical multiplier update and the classical AL function, namely, (4) and (2) with $(\chi, \theta) = (1, 0)$, respectively.

*Organization of the Paper.* Subsection 1.1 provides some basic definitions and notation. Section A presents some necessary background material. Section 2 contains three subsections. The first one describes the main problem of interest and the assumptions made on it. The second one, states an important procedure that is used to generate approximate stationary points of (1). The third one presents the AIDAL method and states its iteration complexity. Section 3 is divided into two subsections and each subsection gives the proof of a key proposition in Section 2. Section 4 presents numerical experiments that demonstrate the efficiency of the AIDAL method. Section 5 gives some concluding remarks. Finally, the end of the paper contains several important technical appendices.

### 1.1 Basic Notations and Definitions

This subsection presents notation and basic definitions used in this paper.

Let $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote the set of nonnegative and positive real numbers, respectively, and let $\mathbb{R}^n$ denote the $n$-dimensional Hilbert space with inner product and associated norm denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively. The smallest positive singular value of a nonzero linear operator $Q : \mathbb{R}^n \to \mathbb{R}^l$ is denoted by $\sigma_Q^+$. For a given closed convex set $X \subset \mathbb{R}^n$, its boundary is denoted by $\partial X$ and the distance of a point $x \in \mathbb{R}^n$ to $X$ is denoted by $\mathrm{dist}_X(x)$. For any $t > 0$, we let

$\log_1^+(t) := \max\{\log t, 1\}$ and $\bar{B}(0,t) := \{z \in \mathbb{R}^n : \|z\| \le t\}$. Using the asymptotic notation $\mathcal{O}$, we denote $\mathcal{O}_1 = \mathcal{O}(\cdot + 1)$.

The domain of a function $h : \mathbb{R}^n \to (-\infty, \infty]$ is the set $\operatorname{dom} h := \{x \in \mathbb{R}^n : h(x) < +\infty\}$. Moreover, $h$ is said to be proper if $\operatorname{dom} h \neq \emptyset$. The set of all lower semi-continuous proper convex functions defined in $\mathbb{R}^n$ is denoted by $\overline{\operatorname{Conv}}\, \mathbb{R}^n$. The $\varepsilon$-subdifferential of a proper function $h : \mathbb{R}^n \to (-\infty, \infty]$ is defined by

$$\partial_\varepsilon h(z) := \{u \in \mathbb{R}^n : h(z') \ge h(z) + \langle u, z' - z\rangle - \varepsilon, \quad \forall z' \in \mathbb{R}^n\} \tag{7}$$

for every $z \in \mathbb{R}^n$. The classical subdifferential, denoted by $\partial h(\cdot)$, corresponds to $\partial_0 h(\cdot)$. The normal cone of a closed convex set $C$ at $z \in C$, denoted by $N_C(z)$, is defined as

$$N_C(z) := \{\xi \in \mathbb{R}^n : \langle \xi, u - z\rangle \le \varepsilon, \quad \forall u \in C\}.$$

If $\psi$ is a real-valued function which is differentiable at $\bar{z} \in \mathbb{R}^n$, then its affine approximation $\ell_\psi(\cdot, \bar{z})$ at $\bar{z}$ is given by

$$\ell_\psi(z; \bar{z}) := \psi(\bar{z}) + \langle \nabla\psi(\bar{z}), z - \bar{z}\rangle \quad \forall z \in \mathbb{R}^n. \tag{8}$$

# 2 Augmented Lagrangian Method

This section contains two subsections. The first one precisely describes the problem of interest and the assumptions underlying it, while the second one presents the AIDAL method and its corresponding iteration complexity.

## 2.1 Problem of Interest

This subsection presents the main problem of interest and the assumptions underlying it.

Our problem of interest is precisely (1) where $f$, $h$, $A$, and $b$ are assumed to satisfy the following assumptions:

(A1) $h \in \overline{\operatorname{Conv}}\, \mathbb{R}^n$ is $K_h$-Lipschitz continuous and $\mathcal{H} := \operatorname{dom} h$ is compact with diameter $D_h := \sup_{u,z\in\mathcal{H}} \|u - z\| < \infty$.

(A2) $f$ is differentiable function on $\mathcal{H}$, and there exists $(m, M) \in \mathbb{R}^2_{++}$ satisfying $m \le M$, such that for every $u, z \in \mathcal{H}$, we have

$$-\frac{m}{2}\|u - z\|^2 \le f(u) - f(z) - \langle \nabla f(z), u - z\rangle \le \frac{M}{2}\|u - z\|, \tag{9}$$

$$\|\nabla f(u) - \nabla f(z)\| \le M\|u - z\|; \tag{10}$$

(A3) there exists $\bar{z} \in \operatorname{int} \mathcal{H}$ such that $A\bar{z} = b$;

(A4) $A \neq 0$, $\mathcal{F} := \{z \in \mathcal{H} : Az = b\} \neq \emptyset$, and $\inf_{z\in\mathbb{R}^n} \phi(z) > -\infty$.

We now make three remarks about the above assumptions. First, it is well-known that (10) implies that $|f(u) - \ell_u(u; z)| \le M\|u - z\|^2/2$ for every $u, z \in \mathcal{H}$ and hence that (9) holds with $m = M$. However, we show that better iteration complexities can be derived when a scalar $m < M$ satisfying (9) is available. Second, (9) implies that the function $f + m\|\cdot\|^2/2$ is convex on $\mathcal{H}$. Finally, since $\mathcal{H}$ is compact by (A1), the image of any continuous $\mathbb{R}^k$-valued function, e.g., $u \mapsto \nabla f(u)$, is bounded.

Under the above assumptions, we now briefly discuss the notion of an approximate stationary point of (1) that the AIDAL method uses. It is well-known that a necessary condition for a point $\bar{z}$ to be a local minimum of (1) is that there exists a multiplier $\bar{p}$ such that

$$0 \in \nabla f(\bar{z}) + \partial h(\bar{z}) + A^* \bar{p}, \quad A\bar{z} = b. \tag{11}$$

In view of this fact, we say that a pair $([\hat{z}, \hat{p}], \hat{v})$ is a $(\hat{\rho}, \hat{\eta})$-approximate stationary point of (1) if it satisfies condition (6), which is clearly a relaxation of (11) for any $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$.

### 2.2 AIDAL Method

This section presents the AIDAL method and its corresponding iteration complexity.

We first state the AIDAL method in Algorithm 2.1. Its main steps are: (i) invoking an ACG algorithm to apply the update in (3); (ii) computing a "refined" pair $([\hat{z}, \hat{p}], \hat{v}) = ([\hat{z}_k, \hat{p}_k], \hat{v}_k)$ satisfying the inclusion and (possibly) the inequality in (6); (iii) performing a novel test to determine the next penalty parameter $c_{k+1}$; and (iv) applying the update in (4).

Some more specific remarks about Algorithm 2.1 are in order. First, its input $z_0$ can be any element in $\mathcal{H}$ and does not necessarily need to be a feasible point, i.e., one satisfying $Az_0 = b$. Second, its steps 1 and 4 are respectively the updates (3) and (4), while its step 3 consists of a test to determine whether the penalty parameter $c_k$ should be increased. In particular, the update for (3) is obtained by applying the ACG algorithm in Algorithm A.1 to the (convex) proximal subproblem

$$\min_{u \in \mathbb{R}^n} \left\{ \lambda \mathcal{L}^\theta_{c_k}(\cdot; p_{k-1}) + \frac{1}{2}\| \cdot - z_{k-1}\|^2 \right\}$$

with an inexactness criterion (see (47)) previously considered by the authors in [9,11,12]. Third, it performs two kinds of iterations: (i) the ones indexed by $k$; and (ii) the ones that are performed by the ACG algorithm every time it is called in its Step 1. To be concise, the former will be referred to as "outer" iterations and the latter as "inner" (ACG) iterations. Finally, the computations in its step 2 are performed to ensure that the its output triple aligns with the notion of a $(\hat{\rho}, \hat{\eta})$-approximate stationary point from the previous subsection (see Lemma 3.1 for more details).

We now present the key properties of the method. To keep the statements concise, we introduce the useful constants

$$\begin{gathered} \bar{d} := \mathrm{dist}_{\partial \mathcal{H}}(\bar{z}), \quad G_f := \sup_{u \in \mathcal{H}} \|\nabla f(u)\|, \quad \phi_* := \inf_{u \in \mathbb{R}^n} \phi(u), \quad \phi^* := \inf_{u \in \mathcal{F}} \phi(u), \\ \tilde{\beta}_\lambda := K_h + G_f + \frac{(1+3\nu)D_h}{\lambda(1-\sigma)}, \quad \beta_\lambda := \tilde{\beta}_\lambda \left(\bar{d} + D_h\right) + \frac{1}{2\lambda}\left(\frac{\sigma D_h}{1-\sigma}\right)^2, \end{gathered} \tag{13}$$

where $(D_h, K_h, \mathcal{H})$, $\bar{z}$, and $\mathcal{F}$ are as in (A1), (A3), and (A4), respectively. Moreover, we let

$$\mathcal{C}_\ell := \left\{ k \in \mathbb{N} : k \geq 2, c_k = c_1 2^{\ell-1} \right\} \tag{14}$$

denote the $\ell^{\text{th}}$ *cycle* of AIDAL method, and for simplicity, if the AIDAL method terminates at iteration $k$ then the indices of the last cycle do not extend past $k$.

The first result, whose proof is the topic of Subsection 3.2, presents some bounds on the quantities $\|p_k\|$ and $\mathcal{L}^\theta_{c_k}(z_k; p_k)$.

**Proposition 2.1.** *Let $\{(c_i, z_i, p_i)\}_{i \geq 1}$ be generated by the AIDAL method. Then, for any $j \geq 1$ and $k \geq 1$, we have*

$$\|p_k\| \leq B_p, \quad \mathcal{L}^\theta_{c_j}(z_j; p_j) - \mathcal{L}^\theta_{c_k}(z_k; p_k) \leq B_{\mathcal{L}}.$$

**Algorithm 2.1:** Accelerated Inexact Dampened Augmented Lagrangian (AIDAL) Method

**Require:** $(\hat{\rho},\hat{\eta}) \in \mathbb{R}^2_{++}$, $(z_0,p_0) \in \mathcal{H} \times A(\mathbb{R}^n)$, $\lambda \in (0, 1/m)$, $(\nu,\sigma,c_1) \in \mathbb{R}^3_{++}$, $[\min\{\sigma,(1-\sigma)\}]^{-1} = \mathcal{O}(1)$, and $(\chi,\theta) \in (0,1)^2$ satisfying

$$\chi \le \frac{\theta^2}{2(1-\theta)(2-\theta)}. \tag{12}$$

**Output :** a pair $([\hat{z},\hat{p}],\hat{v}) \in [\mathcal{H} \times A(\mathbb{R}^n)] \times \mathbb{R}^n$ satisfying (6).

**Function** `AIDAL(`$[f,h,A,b],[\lambda,M],[\nu,\sigma,\chi,\theta],[c_1,z_0,p_0],[\hat{\rho},\hat{\eta}]$`):`
- **for** $k \leftarrow 1,2,...$ **do**
  - `STEP 0` (helper quantities):
  - $L_k \leftarrow \lambda(M + c_k\|A\|^2) + 1$
  - $\psi_s^k(\cdot) \leftarrow \lambda\left[\mathcal{L}^\theta_{c_k}(\cdot;p_{k-1}) - h(\cdot)\right] + \frac{1}{4}\|\cdot - z_{k-1}\|^2$ ▷ See (2) for the definition of $\mathcal{L}^\theta_c(\cdot;\cdot)$
  - $\psi_n^k(\cdot) \leftarrow \lambda h(\cdot) + \frac{1}{4}\|\cdot - z_{k-1}\|^2$
  - `STEP 1` (inexact prox update): ▷ Implement (3)
  - $\tilde{\sigma}_k \leftarrow \min\left\{\frac{\nu}{\sqrt{L_k}},\sigma\right\}$ ▷ Inexactness parameter for ACG
  - $(z_k,v_k,\varepsilon_k) \leftarrow$`ACG(`$[\psi_s^k,\psi_n^k],[L_k-\frac{1}{2},\frac{1}{2}],\tilde{\sigma}_k,z_{k-1}$`)` ▷ See Algorithm A.1
  - `STEP 2` (refinement and termination check):
  - $\tilde{\psi}_s^k(\cdot) \leftarrow \psi_s^k(\cdot) + \frac{1}{4}\|\cdot - z_{k-1}\|^2 - \langle v_k,\cdot\rangle$
  - $\hat{z}_k \leftarrow \operatorname{argmin}_{u\in\mathbb{R}^n}\left\{\ell_{\tilde{\psi}_s^k}(u;z_k) + \frac{L_k}{2}\|u-z_k\|^2 + \lambda h(u)\right\}$
  - $\hat{v}_k \leftarrow \frac{1}{\lambda}\left[(v_k + z_{k-1} - \hat{z}_k) + L_k(z_k - \hat{z}_k) + \nabla\tilde{\psi}_s^k(\hat{z}_k) - \nabla\tilde{\psi}_s^k(z_k)\right]$
  - $\hat{p}_k \leftarrow (1-\theta)p_{k-1} + c_k(A\hat{z}_k - b)$
  - **if** $\|\hat{v}_k\| \le \hat{\rho}$ **and** $\|A\hat{z}_k - b\| \le b$ **then**
    - **return** $([\hat{z}_k,\hat{p}_k],\hat{v}_k)$ ▷ Stop and output
  - `STEP 3` (penalty parameter update): ▷ Compute $c_{k+1}$
  - **if** $\|A(\hat{z}_k - z_k)\| \le \frac{\hat{\eta}}{2}$ **and** $\|A\hat{z}_k - b\| > \hat{\eta}$ **then**
    - $c_{k+1} \leftarrow 2c_k$
  - **else**
    - $c_{k+1} \leftarrow c_k$
  - `STEP 4` (multiplier update): ▷ Implement (4)
  - $p_k \leftarrow (1-\theta)p_{k-1} + \chi c_k(Az_k - b)$

*where*

$$B_p := \frac{\max\{\|p_0\|,\beta_\lambda\}}{\min\{1,\bar{d}\sigma_A^+(1-\theta)\}}, \quad B_{\mathcal{L}} := \phi^* - \phi_* + \frac{1}{\lambda}D_h^2 + \left[\frac{(3-\theta)(1-\theta)}{2\chi c_1}\right]B_p^2. \tag{15}$$

The next result, whose proof is the topic of Subsection 3.3, describes the behavior of $c_k$, a bound on $|\mathcal{C}_\ell|$, the inner iteration complexity of the AIDAL method at each outer iteration, and the output of the method.

**Proposition 2.2.** *Let $\{(c_i,z_i,p_i)\}_{k\ge1}$ be generated by the AIDAL method under a tolerance pair*

$(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$ *and let* $(B_p, B_{\mathcal{L}})$ *be as in Proposition 2.1. Moreover, define the nonnegative scalars*

$$\begin{aligned} \bar{c} = \bar{c}(\hat{\eta}) &:= \frac{4B_p}{\chi\hat{\eta}}, \quad B_\Psi := B_{\mathcal{L}} + \frac{B_p^2}{\chi^2 c_1}, \\ \mathcal{T} = \mathcal{T}(\hat{\rho}, \hat{\eta}) &:= \left\lceil 1 + \left(\frac{2\lambda B_\Psi}{1-\sigma^2}\right) \max\left\{\frac{4\nu^2\|A\|^2}{\hat{\eta}^2}, \frac{1+3\nu}{\lambda^2\hat{\rho}^2}\right\}\right\rceil. \end{aligned} \tag{16}$$

*Then, the following statements hold:*

*(a) if* $c_k \geq \bar{c}$*, then* $c_{k'} = c_k$ *for every* $k' \geq k$*;*

*(b) for every* $\ell \geq 1$*, it holds that* $|\mathcal{C}_\ell| \leq \mathcal{T}$*;*

*(c) for every outer iteration* $k \geq 1$*, the AIDAL method performs at most*

$$\left\lceil 1 + \sqrt{2L_k} \log_1^+ \left(2\sqrt{2}\left[\frac{L_k}{\sigma} + \frac{\sqrt{L_k}}{\nu}\right]\right)\right\rceil$$

*inner (ACG) iterations;*

*(d) the AIDAL method always outputs* $(\hat{\rho}, \hat{\eta})$*-approximate stationary point of (1).*

We now ready to present the total iteration complexity of the method.

**Theorem 2.3.** *Let* $(\bar{c}, \mathcal{T})$ *be as (16) and* $L_1$ *be as in step 0 of the AIDAL method at* $k = 1$*. Then, the AIDAL method stops with a* $(\hat{\rho}, \hat{\eta})$*-approximate stationary point of (1) in*

$$\mathcal{O}_1\left(\mathcal{T}\sqrt{\bar{c}L_1}\log_1^+\left[\frac{\bar{c}L_1}{\sigma} + \frac{\sqrt{\bar{c}}L_1}{\nu}\right]\right) \tag{17}$$

*inner (ACG) iterations.*

*Proof.* In view of Proposition 2.2(a) and (c), the total number of inner (ACG) iterations performed by the method is on the order of

$$\mathcal{O}_1\left(\sum_{\ell=1}^{\lceil \log_2 \bar{c}\rceil}\sum_{j\in\mathcal{C}_\ell}\sqrt{L_j}\log_1^+\left[\frac{L_j}{\nu}\right]\right). \tag{18}$$

To simplify this sum, we first note that if $c_j \in \mathcal{C}_\ell$, then the bound $\lambda \leq 1/m \leq 1/M$ implies that

$$L_j = \lambda\left(M + 2^{\ell-1}c_1\|A\|^2 + \lambda^{-1}\right) \leq \lambda\left(2M + 2^{\ell-1}c_1\|A\|^2\right) \leq 2^\ell L_1. \tag{19}$$

Combining (19) with Proposition 2.2(b), it holds that

$$\begin{aligned} \sum_{\ell=1}^{\lceil \log_2 \bar{c}\rceil}\sum_{j\in\mathcal{C}_\ell}\sqrt{L_j} &\leq \mathcal{T}\sqrt{L_1}\sum_{\ell=1}^{\lceil \log_2 \bar{c}\rceil} 2^{\ell/2} = \mathcal{T}\sqrt{L_1}\cdot\sqrt{2}\left(1+\sqrt{2}\right)\left(2^{\lceil \log_2 \bar{c}\rceil/2} - 1\right) \\ &\leq 4\mathcal{T}\sqrt{L_1}\left(2^{\log_2\sqrt{\bar{c}}}\cdot 2^{1/2}\right) = \mathcal{O}_1\left(\mathcal{T}\sqrt{\bar{c}L_1}\right) \end{aligned} \tag{20}$$

Moreover, denoting $\bar{\ell} = \lceil \log_2 \bar{c}\rceil$, it follows from (19) that

$$\max_{1\leq\ell\leq\lceil \log_2 \bar{c}\rceil}\max_{j\in\mathcal{C}_\ell}\left\{\log_1^+ L_j\right\} = \log_1^+\left[\lambda\left(M + c_{\bar{\ell}}A\|^2\right) + 1\right] = \mathcal{O}_1\left(\log_1^+\left[\bar{c}L_1\right]\right). \tag{21}$$

The complexity bound in (17) now follows from (20), (21), and (18). The conclusion now follows from (17) and Proposition 2.2(d). □

Before ending this section, we give some remarks about the above results. First, Proposition 2.1 states that the sequence of Lagrange multipliers $\{p_k\}_{k\geq 1}$ generated by the AIDAL method is bounded by a constant that is independent of the tolerances $\hat\rho$ and $\hat\eta$. Second, Proposition 2.2(a) states that the number of times that the the penalty constant $c_k$ is doubled during an invocation of the AIDAL method is finite. Third, in terms of only the tolerance pair $(\hat\rho,\hat\eta)$ and the stepsize $\lambda$, Theorem 2.3 states that the AIDAL outputs a $(\hat\rho,\hat\eta)$-approximate stationary point of (1) in

$$\mathcal{O}_1\left(\frac{\lambda^{3/2}}{\hat\eta^{1/2}}\left[\frac{1}{\lambda^2\hat\rho^2}+\frac{1}{\hat\eta^2}\right]\log_1^+\frac{\lambda}{\hat\eta}\right) \tag{22}$$

inner (ACG) iterations. Since the above bound is generally smaller the larger $\lambda$ is, we can infer that the AIDAL method generally performs well when a constant $m \ll M$ is available and $\lambda \approx 1/m$.

# 3 Convergence Analysis of the AIDAL Method

This section establishes the key properties of the AIDAL method and contains three subsections. The first one presents some basic technical results, the second one gives the proof of Proposition 2.1, and the third one gives the proof of Proposition 2.2.

To avoid repetition, we let

$$\{(z_i,p_i,v_i,\varepsilon_i,\hat z_i,\hat p_i,\hat v_i)\}_{i=1}^k,\quad \{(\psi_s^i,\psi_n^i,\tilde\psi_s^i,c_k,\tilde\sigma_k,L_k)\}_{i=1}^k,$$

denote the two main sequence of iterates generated by the AIDAL method up to and including the $k^{\text{th}}$ iteration. Moreover, for every $i\geq 1$ and any $(\chi,\theta)\in\mathbb{R}^2_{++}$, we make use of the following useful constants and shorthand notation

$$\begin{aligned} a_\theta = \theta(1-\theta),\quad b_\theta := (2-\theta)(1-\theta),\quad \alpha_{\chi,\theta} := \frac{(1-2\chi b_\theta)-(1-\theta)^2}{2\chi}, \\ f_i := Az_i - b,\quad \hat f_i := A\hat z_i - b,\quad \Delta p_i = p_i - p_{i-1},\quad \Delta z_i = z_i - z_{i-1} \end{aligned} \tag{23}$$

as well as the key quantities $(B_p, B_{\mathcal{L}})$ and $\mathcal{C}_\ell$ given in (15) and (14), respectively.

## 3.1 Basic Technical Results

This subsection presents some key technical inequalities and relationships that will useful in establishing the results in later subsections. For the sake of brevity, we assume throughout that $k\geq 2$ is an outer iteration index of AIDAL method.

We first present some basic properties about the iterates generated by the method.

**Lemma 3.1.** *Let $i\leq k$ be fixed and define the quantities*

$$\begin{aligned} \delta_i = \frac{1}{\lambda}\varepsilon_i,\quad r_i := v_i + z_{i-1} - z_i, \\ w_i = r_i + L_i(z_i - \hat z_i),\quad \tilde p_i := (1-\theta)p_{i-1} + c_i(Az_i - b). \end{aligned} \tag{24}$$

*Then, the following statements hold:*

*(a)* $f_i = [p_i - (1-\theta)p_{i-1}]/(\chi c_i)$*;*

*(b) if* $i\geq 2$*, then* $\chi(c_i f_i - c_{i-1}f_{i-1}) = \Delta p_i - (1-\theta)\Delta p_{i-1}$*;*

*(c)* $v_i \in \partial_{\varepsilon_i}(\psi_s^i + \psi_n^i)(z_i)$ *and* $\|v_i\|^2 + 2\varepsilon_i \le \tilde{\sigma}_i^2 \|r_i\|^2$;

*(d)* $w_i \in \nabla f(z_i) + \partial_{\delta_i} h(z_i) + A^* \tilde{p}_i$ *and*

$$\|w_i\| \le \left(\frac{1+\nu}{\lambda}\right)\|r_i\|, \quad \delta_i \le \frac{\nu^2}{2\lambda}\|r_i\|^2;$$

*(e)* $\hat{v}_i \in \nabla f(\hat{z}_i) + \partial h(\hat{z}_i) + A^* \hat{p}_i$ *and*

$$\|\hat{v}_i\| \le \left(\frac{1+3\nu}{\lambda}\right)\|r_i\|, \quad \|\hat{f}_i\| \le \|A(\hat{z}_i - z_i)\| + \|f_i\|.$$

*Proof.* (a) This follows from the definition of $f_i$ in (23), and step 4 of the AIDAL method.

(b) This follows from part (a) and the definition of $\Delta p_i$ in (23).

(c) This follows from the ACG call in step 1 of the AIDAL method, Proposition A.1(b) with $(\psi_s, \psi_n) = (\psi_s^i, \psi_n^i)$ and $(z, v, \varepsilon, z^-) = (z_i, v_i, \varepsilon_i, z_{i-1})$, and the definition of $r_i$.

(d) Observe that part (c) and the definition of $\tilde{\psi}_s^i$ imply $0 \in \partial_{\varepsilon_i}(\tilde{\psi}_s^i + \lambda h)(z_i)$. Moreover, notice that assumption (A2) implies $\nabla \tilde{\psi}_s^i$ is $L_k$-Lipschitz continuous. Using the definitions of $\hat{v}_i$, $w_i$, and $\tilde{p}_i$ and Lemma C.5 with $(\psi_s, \psi_n) = (\tilde{\psi}_s^i, \lambda h)$ and $L = M_i$, it now follows that

$$\begin{aligned}
\lambda w_i &= r_i + M_i(z_i - \hat{z}_i) \in r_i + \nabla \tilde{\psi}_s^i(z_i) + \partial_{\varepsilon_i}(\lambda h)(z_i) \\
&= \lambda \left[\nabla f(z_i) + (1-\theta) A^* p_{i-1} + c_i A^*(Az_i - b)\right] + \partial_{\varepsilon_i}(\lambda h)(z_i) \\
&= \lambda \left[\nabla f(z_i) + A^* \tilde{p}_i\right] + \partial_{\varepsilon_i}(\lambda h)(z_i).
\end{aligned}$$

Combining the above relation with the fact that $\lambda^{-1}\partial_{\varepsilon_i}(\lambda h)(z_i) = \partial_{\varepsilon_i/\lambda} h(z_i)$ and the definition of $\delta_i$ yields the desired inclusion. To show the desired bounds on $\|w_i\|$ and $\delta_i$, we first use part (c), the definition of $\sigma_i$, the triangle inequality, and Lemma C.5 as before to obtain

$$\|\lambda w_i\| \le \|r_i\| + \|L_i(z_i - \hat{z}_i)\| \le \|r_i\| + \sqrt{2L_i\varepsilon_i} \le \left(1 + \sqrt{\tilde{\sigma}_i^2 L_i}\right)\|r_i\| \le (1+\nu)\|r_i\|,$$

which clearly implies the desired bound on $\|w_i\|$. On the other hand, using again Lemma C.5 as before and the fact that $\sigma_i \le \nu$ yields $\lambda 2\delta_i = 2\varepsilon_i \le \nu^2 \|r_i\|^2$.

(e) Following the same reasoning as in the proof of part (d), we first observe that $0 \in \partial_{\varepsilon_i}(\tilde{\psi}_s^i + \lambda h)(z_i)$ and $\nabla \tilde{\psi}_s^i$ is $L_k$-Lipschitz continuous. Using the definitions of $\hat{v}_i$ and $\hat{p}_i$ with Lemma C.5 with $(\psi_s, \psi_n) = (\tilde{\psi}_s^i, \lambda h)$ and $L = L_i$, it now follows that

$$\begin{aligned}
\lambda \hat{v}_i &= (v_i + z_{i-1} - \hat{z}_i) + L_i(z_i - \hat{z}_i) + \nabla \tilde{\psi}_s^i(\hat{z}_i) - \nabla \tilde{\psi}_s^i(z_i) \\
&\in (v_i + z_{i-1} - \hat{z}_i) + \nabla \tilde{\psi}_s^i(\hat{z}_i) + \partial(\lambda h)(\hat{z}_i) \\
&= \lambda \left[\nabla f(\hat{z}_i) + (1-\theta) A^* p_{i-1} + c_i A^*(A\hat{z}_i - b)\right] + \partial_{\varepsilon_i}(\lambda h)(\hat{z}_i) \\
&= \lambda \left[\nabla f(\hat{z}_i) + A^* \tilde{p}_i\right] + \partial(\lambda h)(\hat{z}_i).
\end{aligned}$$

which clearly implies the desired inclusion. To show the desired bounds on $\|\hat{v}_i\|$ and $\|\hat{f}_i\|$, we use part (c), the definition of $\sigma_i$, the triangle inequality, and Lemma C.5 as before to obtain

$$\begin{aligned}
\|\lambda \hat{v}_i\| &\le \|r_i\| + (\tilde{\psi}_s^i + 1)\|z_i - \hat{z}_i\| + \|\nabla \tilde{\psi}_s^i(\hat{z}_i) - \nabla \tilde{\psi}_s^i(z_i)\| \le \|r_i\| + (2L_i + 1)\|z_i - \hat{z}_i\| \\
&\le \|r_i\| + (2L_i + 1)\sqrt{\frac{2\varepsilon_i}{L_i}} \le \left(1 + \frac{(2L_i+1)\sqrt{\tilde{\sigma}_i^2}}{\sqrt{L_i}}\right)\|r_i\| \le (1 + 3\sqrt{\tilde{\sigma}_i^2 L_i})\|r_i\| \\
&\le (1+3\nu)\|r_i\|,
\end{aligned}$$

which clearly implies the desired bound on $\|\hat{v}_i\|$. On the other hand, the definitions of $f_i$ and $\hat{f}_i$ and the triangle inequality yield $\|\hat{f}_i\| = \|A(\hat{z}_i - z_i) + f_i\| \le \|A(\hat{z}_i - z_i)\| + \|f_i\|$. □

The next result characterizes the behavior of $\mathcal{L}^\theta_{c_i}(\cdot;\cdot)$ at consecutive iterations of the method.

**Lemma 3.2.** *Let $a_\theta$ and $b_\theta$ be as in (23). Then, the following statements hold:*

*(a) for every $1 \le i \le k$, we have*

$$\mathcal{L}^\theta_{c_i}(z_i;p_i) - \mathcal{L}^\theta_{c_i}(z_i;p_{i-1}) = \frac{b_\theta}{2\chi c_i}\|\Delta p_i\|^2 + \frac{a_\theta}{2\chi c_i}\left(\|p_i\|^2 - \|p_{i-1}\|^2\right); \tag{25}$$

*(b) for every $2 \le i \le k$, we have*

$$\mathcal{L}^\theta_{c_i}(z_i;p_{i-1}) - \mathcal{L}^\theta_{c_i}(z_{i-1};p_{i-1}) \le -\left(\frac{1-\sigma^2}{2\lambda}\right)\|r_i\|^2 - \left(\frac{1-\lambda m}{4\lambda}\right)\|\Delta z_i\|^2 - \frac{c_i}{4}\|A\Delta z_i\|^2. \tag{26}$$

*Proof.* (a) Let $1 \le i \le k$ be fixed. Using the definition of $\mathcal{L}^\theta_c$ in (2), the definitions of $\Delta p_i$ and $f_i$ in (23), and Lemma 3.1(a), we have that

$$\begin{aligned}
\mathcal{L}^\theta_{c_i}(z_i,p_i) - \mathcal{L}^\theta_{c_i}(z_i,p_{i-1}) &= (1-\theta)\langle \Delta p_i, f_i\rangle = \left(\frac{1-\theta}{\chi c_i}\right)\|\Delta p_i\|^2 + \frac{(1-\theta)\theta}{\chi c_i}\langle \Delta p_i, p_{i-1}\rangle \\
&= \left(\frac{1-\theta}{\chi c_i}\right)\|\Delta p_i\|^2 + \frac{(1-\theta)\theta}{\chi c_i}\left(\langle p_i, p_{i-1}\rangle - \|p_{i-1}\|^2\right) \\
&= \left(\frac{1-\theta}{\chi c_i}\right)\|\Delta p_i\|^2 + \frac{(1-\theta)\theta}{\chi c_i}\left(-\frac{1}{2}\|\Delta p_i\|^2 + \frac{1}{2}\|p_i\|^2 - \frac{1}{2}\|p_{i-1}\|^2\right) \\
&= \frac{b_\theta}{2\chi c_i}\|\Delta p_i\|^2 + \frac{a_\theta}{2\chi c_i}\left(\|p_i\|^2 - \|p_{i-1}\|^2\right).
\end{aligned} \tag{27}$$

(b) Let $2 \le i \le k$ be fixed and define the quantities

$$Q_{c_i} := (1-\lambda m)I + c_i\lambda A^*A, \quad \|x\|_{c_i} := \langle x, Q_{c_i}x\rangle \quad \forall x \in \mathbb{R}^n.$$

Moreover, observe that assumption (A2) and the definition of $\mathcal{L}^\theta_c$ imply that $\psi^i_s + \psi^i_n$ is 1-strongly convex with respect to $\|\cdot\|_{c_i}$. As a consequence, using Lemma 3.1(c) and Lemma C.2 with $\|\cdot\|_{\mathcal{Z}} = \|\cdot\|_{c_i}$, $\psi = \psi^i_s + \psi^i_n - \langle v, \cdot\rangle$, and $(\xi, y, \eta) = (1, z, \varepsilon)$, we have that

$$v_i \in \partial_{2\varepsilon_i}\left(\lambda\mathcal{L}^\theta_{c_i}(\cdot;p_{i-1}) + \frac{1}{2}\|\cdot - z_{i-1}\|^2 - \frac{1}{4}\|\cdot - z_i\|^2_{c_i}\right)(z_i).$$

which, in particular, implies that

$$\lambda\mathcal{L}^\theta_{c_i}(z_{i-1};p_{i-1}) - \frac{1}{4}\|\Delta z_i\|^2_{c_i} \ge \lambda\mathcal{L}^\theta_{c_i}(z_i;p_{i-1}) + \frac{1}{2}\|\Delta z_i\|^2 - \langle v_i, \Delta z_i\rangle - 2\varepsilon_i. \tag{28}$$

Using then (28), the definitions of $r_i$ and $\delta_i$ in (23), step 2 of the AIDAL method, and Lemma 3.1(c), we conclude that

$$\begin{aligned}
\mathcal{L}^\theta_{c_i}(z_i,p_{i-1}) - \mathcal{L}^\theta_{c_i}(z_{i-1},p_{i-1}) &\le -\frac{1}{4\lambda}\|\Delta z_i\|^2_{c_i} - \frac{1}{2\lambda}\|\Delta z_i\|^2 + \frac{1}{\lambda}\langle v_i, \Delta z_i\rangle + \frac{2}{\lambda}\varepsilon_i \\
&= -\left(\frac{1-\lambda m}{4\lambda}\right)\|\Delta z_i\|^2 - \frac{c_i}{4}\|A\Delta z_i\|^2 + \frac{1}{2\lambda}\left(\|v_i\|^2 + 2\varepsilon_i - \|r_i\|^2\right) \\
&\le -\left(\frac{1-\lambda m}{4\lambda}\right)\|\Delta z_i\|^2 - \frac{c_i}{4}\|A\Delta z_i\|^2 - \left(\frac{1-\sigma^2}{2\lambda}\right)\|r_i\|^2.
\end{aligned}$$

□

Finally, the below result presents an important bound on the residuals $\|r_i\|$ and $\|\Delta z_i\|$ using a special potential function.

**Proposition 3.3.** *Let $a_\theta$ and $\alpha_{\chi,\theta}$ be as in (23). Then, for every $i \geq 1$ where $c_i = c_{i+1}$, it holds that*

$$\left(\frac{1-\sigma^2}{2\lambda}\right)\|r_{i+1}\|^2 + \left(\frac{1-\lambda m}{4\lambda}\right)\|\Delta z_{i+1}\|^2 \leq \Psi_i^\theta - \Psi_{i+1}^\theta, \tag{29}$$

*where*

$$\Psi_j^\theta := \mathcal{L}_{c_j}^\theta(z_j;p_j) - \frac{a_\theta}{2\chi c_j}\|p_j\|^2 + \frac{\alpha_{\chi,\theta}}{4\chi c_j}\|\Delta p_j\|^2 \quad \forall j \geq 1. \tag{30}$$

*Proof.* Let $i \geq 1$ where $c_i = c_{i+1}$ be fixed, let $b_\theta$ be as in (23), and observe that (12) implies $2\chi b_\theta \leq \theta^2$. Moreover, define

$$\widehat{\Delta} p_{i+1} := \Delta p_{i+1} - (1-\theta)\Delta p_i$$

and observe that Lemma B.2 with $(\tau, a, b) = (2\chi b_\theta, \Delta p_{i+1}, \Delta p_i)$ implies that

$$\frac{1}{\chi}\|\widehat{\Delta} p_{i+1}\|^2 \geq 2b_\theta\|\Delta p_{i+1}\|^2 + \alpha_{\chi,\theta}\left(\|\Delta p_{i+1}\|^2 - \|\Delta p_i\|^2\right). \tag{31}$$

Using Lemma 3.1(b), the fact that $c_{i+1} = c_i$, and (31), we then have

$$\begin{aligned}
\frac{c_{i+1}}{4}\|A\Delta z_{i+1}\|^2 &= \frac{1}{4\chi^2 c_{i+1}}\|\chi c_{i+1} A(z_{i+1} - z_i)\|^2 = \frac{1}{4\chi^2 c_{i+1}}\|\chi(c_{i+1}f_{i+1} - c_i f_i) + \chi(c_i - c_{i+1})f_{i-1}\|^2 \\
&= \frac{1}{4\chi c_{i+1}}\left[\frac{1}{\chi}\|\widehat{\Delta} p_{i+1}\|^2\right] \geq \frac{1}{4\chi c_{i+1}}\left[2b_\theta\|\Delta p_{i+1}\|^2 + \alpha_{\chi,\theta}\left(\|\Delta p_{i+1}\|^2 - \|\Delta p_i\|^2\right)\right] \\
&= \frac{b_\theta}{2\chi c_{i+1}}\|\Delta p_{i+1}\|^2 + \frac{\alpha_{\chi,\theta}}{4\chi c_{i+1}}\left(\|\Delta p_{i+1}\|^2 - \|\Delta p_i\|^2\right).
\end{aligned} \tag{32}$$

Using now (25), (26), (32), and the assumption that $c_i = c_{i+1}$, yields

$$\begin{aligned}
\mathcal{L}_{c_i}^\theta(z_i, p_i) - \mathcal{L}_{c_{i+1}}^\theta(z_{i+1}, p_{i+1}) &= \mathcal{L}_{c_{i+1}}^\theta(z_i, p_i) - \mathcal{L}_{c_{i+1}}^\theta(z_{i+1}, p_{i+1}) \\
&\geq \left(\frac{1-\sigma^2}{2\lambda}\right)\|r_{i+1}\|^2 + \left(\frac{1-\lambda m}{4\lambda}\right)\|\Delta z_{i+1}\|^2 + \frac{a_\theta}{2\chi c_{i+1}}\left(\|p_i\|^2 - \|p_{i+1}\|^2\right) \\
&\quad - \frac{b_\theta}{2\chi c_i}\|\Delta p_{i+1}\|^2 + \frac{c_{i+1}}{4}\|A\Delta z_{i+1}\|^2 \\
&\geq \left(\frac{1-\sigma^2}{2\lambda}\right)\|r_{i+1}\|^2 + \left(\frac{1-\lambda m}{4\lambda}\right)\|\Delta z_{i+1}\|^2 + \frac{a_\theta}{2\chi c_{i+1}}\left(\|p_i\|^2 - \|p_{i+1}\|^2\right) \\
&\quad + \frac{\alpha_{\chi,\theta}}{4\chi c_{i+1}}\left(\|\Delta p_{i+1}\|^2 - \|\Delta p_i\|^2\right).
\end{aligned} \tag{33}$$

The conclusion now follows from combining (33) with the definition of $\Psi_j^\theta$ in (30). □

## 3.2 Proof of Proposition 2.1

This subsection gives the proof of Proposition 2.1.

We first focus on proving that $\|p_k\| \leq B_p$ for every $k \geq 0$. The first result presents some key properties about some refined iterates.

**Lemma 3.4.** *Let* $(\tilde{p}_i, w_i)$ *be as in* (24) *and define*

$$\xi_i := w_i - \nabla f(z_i) - A^* \tilde{p}_i \quad \forall i \geq 0. \tag{34}$$

*Then, the following statements hold for every* $k \geq 1$*:*

*(a) we have that*

$$\xi_k \in \partial_{\delta_k} h(z_k), \quad \|r_k\| \leq \frac{D_h}{1-\sigma}, \quad \|w_k\| \leq \frac{(1+3\nu)D_h}{\lambda(1-\sigma)}, \quad \delta_k \leq \frac{1}{2\lambda}\left(\frac{\nu D_h}{1-\sigma}\right)^2.$$

*(b) we have that*

$$\|\tilde{p}_k\| \leq \frac{1}{\sigma_A^+}\left[\|\xi_k\| + \|\nabla f(z_k)\| + \frac{(1+\nu)D_h}{\lambda(1-\sigma)}\right].$$

*Proof.* (a) The inclusion follows immediately from Lemma 3.1(d) and the definition of $\xi_k$. For the desired bounds, we use the definition of $r_k$, Lemma 3.1(c), the fact that $\sigma_i \leq \bar{\sigma}^2$ and $z_k, z_{k-1} \in \mathcal{H}$, and the reverse triangle inequality, to first obtain

$$\|r_k\| - D_h \leq \|r_k\| - \|z_k - z_{k-1}\| \leq \|v_k\| \leq \tilde{\sigma}_i \|r_k\| \leq \sigma \|r_k\|.$$

A re-arrangement of the above inequality yields the bound on $\|r_k\|$. The bounds on $\|w_k\|$ and $\delta_k$ now follow from the previous bound on $\|r_k\|$ and Lemma 3.1(d).

(b) Using the definition of $\xi_k$, the triangle inequality, part (b), and Lemma B.1 with $S = A^*$ and $u = \tilde{p}_k$ yields

$$\|\tilde{p}_k\| \leq \frac{1}{\sigma_A^+}\|A^*\tilde{p}_k\| = \frac{1}{\sigma_A^+}\|w_k - \nabla f(z_k) - \xi_k\| \leq \frac{1}{\sigma_A^+}\left[\|\xi_k\| + \|\nabla f(z_k)\| + \frac{(1+3\nu)D_h}{\lambda(1-\bar{\sigma})}\right].$$

□

The next result establishes several technical inequalities.

**Lemma 3.5.** *Let* $(\beta_\lambda, \bar{d})$ *be as in* (13)*. Then, the following statements hold for any* $(\chi, \theta) \in (0,1)^2$ *and* $k \geq 1$*:*

*(a)* $\|p_k\| \leq \chi \|\tilde{p}_k\| + (1-\chi)(1-\theta)\|p_{k-1}\|$;

*(b)* $c_k^{-1}\|\tilde{p}_k\|^2 + \bar{d}\sigma_A^+ \|\tilde{p}_k\| \leq c_k^{-1}(1-\theta)\langle \tilde{p}_k, p_{k-1}\rangle + \beta_\lambda$.

*Proof.* (a) Using the definitions of $p_k$ and $\tilde{p}_k$ with the triangle inequality yields

$$\|p_k\| = \|\chi \tilde{p}_k + (1-\chi)(1-\theta)p_{k-1}\| \leq \chi\|\tilde{p}_k\| + (1-\chi)(1-\theta)\|p_{k-1}\|.$$

(b) Let $\xi_k$, $(G_f, \bar{d})$, and $D_h$ be as in (34), (13), and assumption (A1), respectively. Using Lemma 3.4(a), the definition of $\bar{d}$, and Lemma C.4 with $(\psi, K_\psi, D_\psi) = (h, K_h, D_h)$ and $(y, \bar{y}, \varepsilon) = (\hat{z}_k, \bar{z}, \delta_k)$, we have that

$$\bar{d}\|\xi_k\| \leq (\bar{d} + D_h)K_h + \langle \xi_k, z_k - \bar{z}\rangle + \frac{1}{2\lambda}\left(\frac{\sigma D_h}{1-\sigma}\right)^2. \tag{35}$$

Moreover, the definitions of $\tilde{p}_k$ and $\xi_k$, the fact that $z_k, \bar{z} \in \mathcal{H}$ and $A\bar{z} = b$, and the Cauchy-Schwarz inequality imply that

$$\begin{aligned}\langle \xi_k, z_k - \bar{z} \rangle &= \langle w_k - \nabla f(z_k) - A^* \tilde{p}_k, z_k - \bar{z} \rangle \\ &\le (\|w_k\| + \|\nabla f(z_k)\|) \|z_k - \bar{z}\| - \langle \tilde{p}_k, Az_k - b \rangle \\ &\le \left[ \frac{(1+3\nu)D_h}{\lambda(1-\sigma)} + G_f \right] D_h + \left( \frac{1-\theta}{c_k} \right) \langle \tilde{p}_k, p_{k-1} \rangle - \frac{1}{c_k} \|\tilde{p}_k\|^2. \end{aligned} \tag{36}$$

Using Lemma 3.4(b), (35), (36), and the definitions of $\tilde{\beta}_\lambda$ and $\beta_\lambda$ in (13), we thus conclude that

$$\begin{aligned}\frac{1}{c_k}\|\tilde{p}_k\|^2 + \bar{d}\sigma_A^+ \|\tilde{p}_k\| &\le \frac{1}{c_k}\|\tilde{p}_k\|^2 + \bar{d}\|\xi_k\| + \left[ G_f + \frac{(1+3\nu)D_h}{\lambda(1-\sigma)} \right] \bar{d} \\ &\le \langle \xi_k, z_k - \bar{z} \rangle + \frac{1}{2\lambda} \left( \frac{\sigma D_h}{1-\sigma} \right)^2 + \tilde{\beta}_\lambda \bar{d} \le \left( \frac{1-\theta}{c_k} \right) \langle \tilde{p}_k, p_{k-1} \rangle + \beta_\lambda. \end{aligned}$$

□

We now are now ready to show the desired bound on $\|p_k\|$.

**Proposition 3.6.** *For every $k \ge 0$, it holds that $\|p_k\| \le B_p$.*

*Proof.* We proceed by induction on $k$. Since $B_p \ge \|p_0\|$, the desired bound trivially holds for $k = 0$. Assume now that $\|p_k\| \le B_p$ holds for some $k \ge 0$. If $\|\tilde{p}_{k+1}\| = 0$, then clearly

$$\|p_{k+1}\| \le \chi \|\tilde{p}_{k+1}\| + (1-\chi)(1-\theta)\|p_k\| = (1-\chi)(1-\theta)B_p \le B_p,$$

so suppose that $\|\tilde{p}_{k+1}\| > 0$. Using Lemma 3.5(b), the Cauchy-Schwarz inequality, and the induction hypothesis we have that

$$\begin{aligned}\left[ \bar{d} + \frac{1}{c_{k+1}\sigma_A^+} \|\tilde{p}_{k+1}\| \right] \|\tilde{p}_{k+1}\| &= \frac{1}{\sigma_A^+} \left[ \frac{1}{c_{k+1}} \|\tilde{p}_{k+1}\|^2 + \bar{d}\sigma_A^+ \|\tilde{p}_{k+1}\| \right] \le \frac{1}{\sigma_A^+} \left[ \left( \frac{1-\theta}{c_{k+1}} \right) \langle \tilde{p}_k, p_{k-1} \rangle + \beta_\lambda \right] \\ &= \frac{1}{c_{k+1}\sigma_A^+} \left[ c_{k+1}\beta_\lambda + (1-\theta)\|p_k\| \cdot \|\tilde{p}_{k+1}\| \right] \le \left[ \bar{d} + \frac{1}{c_{k+1}\sigma_A^+} \|\tilde{p}_{k+1}\| \right] B_p. \end{aligned}$$

and hence that $\|\tilde{p}_{k+1}\| \le B_p$. Combining this bound with the induction hypothesis and part (a), we finally conclude that

$$\|p_{k+1}\| \le \chi \|\tilde{p}_{k+1}\| + (1-\chi)(1-\theta)\|p_k\| \le B_p.$$

□

Using the above result, we now show the desired bound on $\mathcal{L}_{c_k}(z_k; p_k)$.

**Proposition 3.7.** *Let $(\phi_*, \phi^*)$ and $D_h$ be as in (13) and assumption (A1), respectively. Then, for every $j \ge 1$ and $k \ge 1$, it holds that*

$$\mathcal{L}_{c_k}^\theta(z_k; p_k) \ge \phi_* - \frac{(1-\theta)^2}{2c_k} B_p^2, \quad \mathcal{L}_{c_j}^\theta(z_j; p_j) \le \phi^* + \frac{D_h^2}{\lambda} + \left[ \frac{a_\theta + 4b_\theta}{2\chi c_j} \right] B_p^2, \tag{37}$$

*and, as a consequence, we have that*

$$\mathcal{L}_{c_j}^\theta(z_j; p_j) - \mathcal{L}_{c_k}^\theta(z_k; p_k) \le B_{\mathcal{L}}. \tag{38}$$

*Proof.* Let $j \geq 1$ and $k \geq 1$ be fixed. Using Lemma 3.6, the fact that $c_i \leq c_{i+1}$, the definitions of $\mathcal{L}_c^\theta$, $\phi_*$, and $B_p$, we have

$$\begin{aligned}
\mathcal{L}_{c_k}^\theta(z_k; p_k) &= \phi(z_k) + (1-\theta)\left\langle p_k, Az_k - b\right\rangle + \frac{c_k}{2}\|Az_k - b\|^2 \\
&\geq \phi_* + \frac{1}{2}\left\|\left(\frac{1-\theta}{\sqrt{c_k}}\right)p_k + \sqrt{c_k}(Az_k - b)\right\|^2 - \frac{(1-\theta)^2}{2c_k}\|p_k\|^2 \\
&\geq \phi_* - \frac{(1-\theta)^2}{2c_k}\|p_k\|^2 \geq \phi_* - \frac{(1-\theta)^2}{2c_k}B_p^2,
\end{aligned}$$

which is the desired lower bound in (37). For the upper bound, let $u \in \mathcal{F}$ be fixed (see assumption (A4)) and use Lemma 3.1(c) and Lemma C.1 with $\psi = \mathcal{L}_{c_j}^\theta(\cdot; p_{j-1})$, $s = 1$, and $(z^-, z, v, \varepsilon) = (z_{j-1}, z_j, v_j, \varepsilon_j)$, to obtain

$$\lambda\mathcal{L}_{c_j}^\theta(z_j; p_{j-1}) + \left(\frac{1-2\sigma^2}{2}\right)\|r_j\|^2 \leq \lambda\mathcal{L}_{c_j}^\theta(u; p_{j-1}) + \|u - z_{j-1}\|^2 = \lambda\phi(u) + D_h^2,$$

and hence that $\mathcal{L}_{c_j}^\theta(z_j; p_{j-1}) \leq \phi^* + D_h^2/\lambda$. Combining the previous bound with Lemma 3.6, Lemma 3.2(a), and the relation $(a+b)^2 \leq 2a^2 + 2b^2$ for every $a, b \in \mathbb{R}$, we now conclude that

$$\begin{aligned}
\mathcal{L}_{c_j}^\theta(z_j; p_j) &\leq \mathcal{L}_{c_j}^\theta(z_j; p_{j-1}) + \frac{b_\theta}{2\chi c_j}\|p_j - p_{j-1}\|^2 + \frac{a_\theta}{2\chi c_j}\|p_j\|^2 \\
&\leq \phi^* + \frac{D_h^2}{\lambda} + \left[\frac{a_\theta + 4b_\theta}{2\chi c_j}\right]B_p^2,
\end{aligned}$$

which is the desired upper bound in (37). To show the last bound, we use the fact that $\{c_k\}_{k\geq 1}$ is a nondecreasing sequence and $\chi \in (0,1)$ to obtain

$$\frac{(1-\theta)^2}{2c_k} + \frac{a_\theta + 4b_\theta}{2\chi c_j} \leq \frac{(1-\theta)^2 + a_\theta + 4b_\theta}{2\chi c_1} = \frac{(3-\theta)(1-\theta)}{2\chi c_1}.$$

The last bound now follows from (37), the above bound, and the definition of $B_{\mathcal{L}}$. □

Combining the previous results easily gives Proposition 2.1. For completeness, we state this explicitly below.

*Proof of Proposition 2.1.* The bound on $\|p_k\|$ follows immediately from Lemma 3.6 whereas the other bound follows from immediately from Lemma 3.7. □

### 3.3 Proof of Proposition 2.2

This subsection gives the proof of Proposition 2.2.

We first give some key bounds involving the residuals $\|r_i\|$, $\|\hat{v}_i\|$, and $\|A(\hat{z}_i - z_i)\|$.

**Lemma 3.8.** *Let $B_\Psi$ be as in (16). Then, for every $\ell \geq 1$ and $(k^-, k^+)$ satisfying*

$$k^- \geq \inf \mathcal{C}_\ell, \quad k_+ \leq \sup \mathcal{C}_\ell, \quad k_- < k_+, \tag{39}$$

*there exists $i \in \{k^- + 1, ..., k^+\}$ such that*

$$\|r_i\|^2 \leq \frac{B_r}{k^+ - k^-}, \quad \|\hat{v}_i\| \leq \left(\frac{1+3\nu}{\lambda}\right)\sqrt{\frac{B_r}{k^+ - k^-}}, \quad \|A(\hat{z}_i - z_i)\| \leq \nu\|A\|\sqrt{\frac{B_r}{k^+ - k^-}}, \tag{40}$$

*where $B_r := 2\lambda B_\Psi/(1-\sigma^2)$.*

*Proof.* Let $\ell \geq 1$ and $(k^-, k^+)$ satisfying (39) be fixed. Moreover, let $\Psi_j^\theta$ be as in (30) and denote $\tilde{c}_\ell = c_{k^-} = c_{k^+}$. We first show the required bound on $\|r_i\|$. Summing (29) from $i = k^-$ to $(k^+ - 1)$ and using Proposition 2.1, the fact that $\{c_k\}_{k\geq 1}$ is a nondecreasing sequence, the bound $a_\theta + \alpha_{\chi,\theta} \leq 2\chi^{-1}$, and the relation $(a+b)^2 \leq 2a^2 + 2b^2$ for every $a, b \in \mathbb{R}$, it holds that

$$
\begin{aligned}
&(k^+ - k^-) \min_{k^-+1\leq i\leq k^+} \|r_i\|^2 \\
&\leq \sum_{i=k^-+1}^{k^+} \|r_i\|^2 \leq \left(\frac{2\lambda}{1-\sigma^2}\right)\left(\Psi_{k^-}^\theta - \Psi_{k^+}^\theta\right) \\
&\leq \left(\frac{2\lambda}{1-\sigma^2}\right)\left[\mathcal{L}_{\tilde{c}_\ell}^\theta(z_{k^-}; p_{k^-}) - \mathcal{L}_{\tilde{c}_\ell}^\theta(z_{k^+}; p_{k^+}) + \frac{\alpha_{\chi,\theta}}{4\chi\tilde{c}_\ell}\|\Delta p_{k^-}\|^2 + \frac{\alpha_\theta}{2\chi\tilde{c}_\ell}\|p_{k^+}\|^2\right] \\
&\leq \left(\frac{2\lambda}{1-\sigma^2}\right)\left[B_{\mathcal{L}} + \frac{\alpha_{\chi,\theta}}{2\chi c_1}\left(\|p_{k^-}\|^2 + \|p_{k^--1}\|^2\right) + \frac{\alpha_\theta}{2\chi c_1}B_p^2\right] \\
&\leq \left(\frac{2\lambda}{1-\sigma^2}\right)\left[B_{\mathcal{L}} + \left(\frac{a_\theta + \alpha_{\chi,\theta}}{2\chi c_1}\right)B_p^2\right] \leq \left(\frac{2\lambda}{1-\sigma^2}\right)\left[B_{\mathcal{L}} + \frac{B_p^2}{\chi^2 c_1}\right] = \frac{2\lambda B_\Psi}{1-\sigma^2},
\end{aligned}
\tag{41}
$$

which implies the existence of some $i \in \{k^- + 1, ..., k^+\}$ satisfying the bound on $\|r_i\|$ in (40). To show the other two bounds, we use Lemma 3.1(d), the fact that $L_j \geq 1$, and Lemma C.5 with $(\psi_s, \psi_n) = (\tilde{\psi}_s^j, \lambda h)$ and $L = M_j$ to obtain

$$
\|\hat{v}_j\| \leq \left(\frac{1+3\nu}{\lambda}\right)\|r_j\|, \quad \|A(\hat{z}_j - z_j)\| \leq \|A\|\sqrt{\frac{2\varepsilon}{L_j}} \leq \frac{\tilde{\sigma}_j\|A\| \cdot \|r_j\|}{\sqrt{L_j}} \leq \nu\|A\| \cdot \|r_j\|,
\tag{42}
$$

for every $j \geq 2$. The result now follows from combining (42) with the previous bound on $\|r_i\|$ at $j = i$. ☐

We are now give the proof of Proposition 2.2.

*Proof of Proposition 2.2.* (a) Let $k \geq 1$ and $k' > k$ be arbitrary and suppose that $c_k \geq \bar{c}$. Using Proposition 2.1, Lemma 3.1(b), and the definition of $\bar{c}$, we first have that

$$
\|f_{k'}\| = \frac{1}{\chi c_{k'}}\|p_{k'} - (1-\theta)p_{k'-1}\| \leq \frac{2B_p}{\chi\bar{c}} \leq \frac{\hat{\eta}}{2}.
\tag{43}
$$

Now, for the sake of contradiction suppose that $k'' > k$ is the first index where $c_{k''} \neq c_k$. In view of the fact that $\{c_k\}_{k\geq 1}$ is nondecreasing, Lemma 3.1(c), and step 3 of the AIDAL method, it must be that $\|Az_{k''} - b\| > \hat{\eta}$ and $\|A(\hat{z}_{k''} - z_{k''})\| \leq \hat{\eta}/2$. However, in view of (43) at $k' = k''$ and Lemma 3.1(e), this is impossible and hence $c_{k'} = c_k$ for every $k' \geq k$.

(b) Let $\ell \geq 1$ be fixed and define $k^- := \inf \mathcal{C}_\ell$ and $k^+ := k^- + \mathcal{T} - 1$. If $c_{k^+} > c_{k^-}$, then clearly we must have $|\mathcal{C}_\ell| \leq k^+ - k^- = \mathcal{T}$. On the other hand, suppose that $c_{k^-} = c_{k^+}$. Then $k^-, k^+ \in \mathcal{C}_\ell$ and, from Lemma 3.8 , there exists $i \in \{k^- + 1, ..., k^+\}$ such that

$$
\|\hat{v}_i\| \leq \left(\frac{1+2\nu}{\lambda}\right)\sqrt{\frac{2\lambda B_\Psi}{(1-\sigma^2)(\mathcal{T}-1)}} \leq \hat{\rho}, \quad \|A(\hat{z}_i - z_i)\| \leq \nu\|A\|\sqrt{\frac{2\lambda B_\Psi}{(1-\sigma^2)(\mathcal{T}-1)}} \leq \frac{\hat{\eta}}{2}.
$$

If $\|A\hat{z}_i - b\| \leq \hat{\eta}$, then the above bound on $\|\hat{v}_i\|$ implies that AIDAL method must stop in step 2 at iteration $i$. Otherwise, we have $\|A\hat{z}_i - b\| > \hat{\eta}$ and the above bound on $\|A(\hat{z}_i - z_i)\|$ together with step 3 of the AIDAL method imply that that $c_{i+1} = 2c_i$. Since $c_{k^-} = c_{k^+}$, it must be that $i = k^+$ and hence we have that $|\mathcal{C}_\ell| = k^+ - k^- + 1 = \mathcal{T}$.

(c) This follows from assumption (A2), step 1 of the AIDAL method, Proposition A.1(b), and the fact that at the $k^{\text{th}}$ iteration we have

$$\left(\frac{1+\tilde{\sigma}_k}{\tilde{\sigma}_k}\right)\sqrt{2L_k} \leq \frac{2\sqrt{2L_k}}{\tilde{\sigma}_k} = \frac{2\sqrt{2L_k}}{\min\{\sigma L_k^{-1/2}, \nu\}} \leq 2\sqrt{2}\left(\frac{L_k}{\sigma} + \frac{\sqrt{L_k}}{\nu}\right).$$

(d) This follows from the inclusion in Lemma 3.1(e) and the termination condition in step 3 of the AIDAL method. □

# 4 Numerical Experiments

This section examines the performance of the AIDAL method for solving problems of the form given in (1). It contains three subsections which each contain one of the following problem classes: (i) a class of linearly-constrained quadratic programming problems considered in [9]; (ii) the sparse principal component analysis (PCA) problem in [5]; and (iii) a class of linearly-constrained quadratic matrix problems considered in [10, 11].

Before proceeding, we first describe some of the implementation details of our experiments. To start, we discuss the three implementations of the AIDAL method in Algorithm 2.1, labeled ADL0–ADL2, considered in this section. First, each implementation uses $\lambda = 1/(2m)$, $\bar{\sigma} = 0.3$, $p_0 = 0$, $c_1 = \max\{1, M/\|A\|^2\}$, and $\nu$ large enough so that $\sigma = \bar{\sigma}$ for every outer iteration of the method. Second, the implementations only differ in their choice of $(\chi, \theta)$ which we summarize in Table 4.1. Note that AIDAL0 uses a parameter pair that does not satisfy (12), but works well in practice.

| **Name** | ADL0 | ADL1 | ADL2 |
|---|---|---|---|
| $\theta$ | 0 | 0.500 | 0.764 |
| $\chi$ | 1 | 0.167 | 1.000 |

**Table 4.1:** AIDAL Parameters for the ADL0–ADL2 variants.

Besides the above AIDAL implementations, we also consider four other methods as benchmarks. The first one, named iALM, is an implementation of the inexact proximal augmented Lagrangian method of [15] in which: (i) its key parameters are

$$\sigma = 5, \quad \beta_0 = \max\left\{1, \frac{\max\{m, M\}}{\|A\|^2}\right\}, \quad w_0 = 1, \quad \boldsymbol{y}^0 = 0, \quad \gamma_k = \frac{(\log 2)\,\|Ax^1\|}{(k+1)\,[\log(k+2)]^2},$$

for every $k \geq 1$; and (ii) the starting point given to the $k^{\text{th}}$ APG call is set to be $\boldsymbol{x}^{k-1}$, which is the prox center for the $k^{\text{th}}$ prox subproblem. The second one, named IPL, is an implementation of the inexact proximal augmented Lagrangian method of [11, Section 5] where: (i) $c_k$ is doubled in its step 4 rather than quintupled; and (ii) $\sigma = 0.3$. The third one, named QP, is a practical modification of the quadratic penalty method of [9] in which: (i) each ACG subproblem in step 1 of the AIPP method is stopped when the condition

$$\|u_j\| + 2\eta_j \leq \sigma\|x_0 - x_j + u_j\|^2$$

holds; and (ii) it uses the parameters $\sigma = 0.3$ and $c = \max\{1, M/\|A\|^2\}$. The fourth and last one, named RQP, is an instance of the relaxed quadratic penalty method of [10] in which: (i) it uses the AIPPv2 variant described in [10, Section 6] with the parameters $(\theta, \tau) = (4, 10[\lambda M + 1])$ and $\lambda = 1/m$; and (ii) it uses the initial penalty parameter $c_1 = \max\{1, M/\|A\|^2\}$.

It should also be mentioned that every method except the iALM replaces its ACG prox subproblem solver by a more practical FISTA variant whose key iterates are as described in [23] and whose main stepsize parameter is adaptively estimated by a line search subroutine described in [8, Algorithm 5.2.1].

For a linear operator $A$, a proper lower semicontinuous convex function $h$, a function $f$ satisfying assumptions (A2)–(A4), tolerance pair $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$, and an initial point $z_0 \in \operatorname{dom} h$, each of the methods of this section seeks a pair $([\hat{z}, \hat{p}], \hat{v})$ satisfying

$$\hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}) + A^* \hat{p}, \quad \frac{\|\hat{v}\|}{\|\nabla f(z_0)\| + 1} \le \hat{\rho}, \quad \frac{\|A\hat{z} - b\|}{\|Az_0 - b\| + 1} \le \hat{\eta}. \tag{44}$$

In particular, all experiments consider a tolerance pair of $(\hat{\rho}, \hat{\eta}) = (10^{-3}, 10^{-3})$.

All of experiments are implemented in MATLAB 2020b and are run on Linux 64-bit machines each containing Xeon E5520 processors and at least 8 GB of memory. Furthermore, the bold numbers in each of the tables of this section indicate the method that performed the most efficiently for a given metric, e.g., runtime or (innermost) iteration count. Finally, it is worth mentioning that the code for replicating these experiments is freely available online[1].

## 4.1 Linearly-Constrained Quadratic Programming

Given a pair of dimensions $(l, n) \in \mathbb{N}^2$, scalar pair $(\alpha_1, \alpha_2) \in \mathbb{R}^2_{++}$, matrices $A, B, C \in \mathbb{R}^{l \times n}$, positive diagonal matrix $D \in \mathbb{R}^{n \times n}$, and vector pair $(b, d) \in \mathbb{R}^l \times \mathbb{R}^l$, this subsection considers the following linearly-constrained quadratic programming (LCQP) problem:

$$\begin{aligned} \min_z \ & \frac{\alpha_1}{2}\|Cz - d\|^2 - \frac{\alpha_2}{2}\|DBz\|^2 \\ \text{s.t. } & Az = b, \quad z \in \Delta_n, \end{aligned}$$

where $\Delta_n = \{z \in \mathbb{R}^n_+ : \sum_{i=1}^n z_i = 1\}$ denotes the $n$-dimensional simplex.

We now describe the experiment parameters for the instances considered. First, the dimensions are set to $(l, n) = (10, 50)$ and *all* of the entries in $A$, $B$, and $C$ are nonzero. Second, the entries of $A, B, C, b$, and $d$ (resp., $D$) are generated by sampling from the uniform distribution $\mathcal{U}[0,1]$ (resp., $\mathcal{U}[1, 1000]$). Third, the initial starting point $z_0$ is generated by sampling a random vector $\tilde{z}_0$ from $\mathcal{U}^2[0,1]$ and setting $z_0 = \tilde{z}_0/\|\tilde{z}_0\|$. Fourth, using the well-known fact that $\|z\| \le 1$ for every $z \in \Delta_n$, the auxiliary parameters for the iALM are

$$B_i = \|a_i\|, \quad L_i = 0, \quad \rho_i = 0 \quad \forall i \ge 1,$$

where $a_i$ is the $i^{\text{th}}$ row of $A$. Finally, the composite form of the problem is

$$f(z) = \frac{\alpha_1}{2}\|Cz - d\|^2 - \frac{\alpha_2}{2}\|DBz\|^2, \quad h(z) = \delta_{\Delta_n}(z),$$

and each problem instance uses a scalar pair $(\alpha_1, \alpha_2) \in \mathbb{R}^2_{++}$ so that $m = -\lambda_{\min}(\nabla^2 f) = 10$ and $M = \lambda_{\max}(\nabla^2 f)$ is a particular value given in the table below.

We now present the numerical results for this set of problem instances in Tables 4.2 and 4.3.

[1] See https://github.com/wwkong/nc_opt/tree/master/tests/papers/aidal.

| $M$ | Iteration Count | | | | | | |
|---|---|---|---|---|---|---|---|
| | ADL0 | ADL1 | ADL2 | iALM | IPL | QP | RQP |
| $10^2$ | 874 | 3282 | 2181 | 15646 | 1157 | 9602 | **551** |
| $10^3$ | 327 | 836 | 613 | 9316 | **283** | 5107 | 598 |
| $10^4$ | 431 | 1269 | 1285 | 11118 | **295** | 2260 | 780 |
| $10^5$ | 1301 | 4069 | 4031 | 16850 | 935 | 5070 | **777** |
| $10^6$ | 9967 | 35615 | 33181 | 30476 | 6940 | 8582 | **667** |

**Table 4.2:** (Innermost) Iteration Counts for LCQP problems.

| $M$ | Runtime (seconds) | | | | | | |
|---|---|---|---|---|---|---|---|
| | ADL0 | ADL1 | ADL2 | iALM | IPL | QP | RQP |
| $10^2$ | 4.44 | 13.71 | 7.50 | 46.17 | 4.81 | 33.59 | **2.16** |
| $10^3$ | 1.35 | 2.88 | 2.48 | 31.74 | **1.24** | 19.74 | 2.52 |
| $10^4$ | 1.69 | 4.67 | 5.11 | 36.25 | **1.16** | 9.18 | 3.14 |
| $10^5$ | 5.24 | 15.52 | 14.27 | 48.74 | **2.81** | 19.72 | 3.19 |
| $10^6$ | 39.42 | 128.41 | 116.43 | 86.18 | 23.04 | 29.79 | **2.45** |

**Table 4.3:** Runtimes for LCQP problems.

## 4.2 Sparse PCA

Given integer $k$, positive scalar pair $(\nu, b) \in \mathbb{R}^2_{++}$, and matrix $\Sigma \in S^n_+$, this subsection considers the following sparse principal component analysis (SPCA) problem:

$$\begin{aligned} \min_{\Pi,\Phi} \ & \langle \Sigma, \Pi \rangle_F + \sum_{i,j=1}^{n} q_\nu(\Phi_{ij}) + \nu \sum_{i,j=1}^{n} |\Phi_{ij}| \\ \text{s.t. } & \Pi - \Phi = 0, \quad (\Pi, \Phi) \in \mathcal{F}^k \times \mathbb{R}^{n \times n}, \end{aligned}$$

where $\mathcal{F}^k = \{z \in S^n_+ : 0 \preceq z \preceq I, \operatorname{tr} M = k\}$ denotes the $k$–Fantope and $q_\nu$ is the minimax concave penalty (MCP) function given by

$$q_\nu(t) := \begin{cases} -t^2/(2b), & \text{if } |t| \le b\nu, \\ b\nu^2/2 - \nu|t|, & \text{if } |t| > b\nu, \end{cases} \qquad \forall t \in \mathbb{R}.$$

Note that the effective domain of this problem is unbounded, and hence, only the QP method is guaranteed to converge to an approximate stationary point in general.

We now describe the experiment parameters for the instances considered. First, the scalar parameters are chosen to be $(\nu, b) = (100, 100, 0.1)$. Second, the matrix $\Sigma$ is generated according to an eigenvalue decomposition $\Sigma = P\Lambda P^T$, based on a parameter pair $(s, k)$, where $k$ is as in the problem description and $s$ is a positive integer. In particular, we choose $\Lambda = (100, 1, ..., 1)$, the first column of $P$ to be a sparse vector whose first $s$ entries are $1/\sqrt{s}$, and the other entries of $P$ to be sampled randomly from the standard Gaussian distribution. Third, the initial starting point is $(\Pi_0, \Phi_0) = (D_k, 0)$ where $D_k$ is a diagonal matrix whose first $k$ entries are 1 and whose remaining entries are 0. Fourth, the curvature parameters for each problem instance are $m = M = 1/b$ and $k$ is fixed at $k = 1$. Fifth, for the iALM, we make the following parameter choices based on a relaxed (but unverified) assumption that its generated iterates lie in $\mathcal{F}_k \times \mathcal{F}_k$:

$$B_i = 1, \quad L_i = 0, \quad \rho_i = 0 \quad \forall i \ge 1.$$

Sixth, the composite form of the problem is

$$f(\Pi,\Phi)=\langle\Sigma,\Pi\rangle_F+\sum_{i,j=1}^{n} q_\nu(\Phi_{ij}),\quad h(\Pi,\Phi)=\delta_{\mathcal{F}^k}(\Pi)+\nu\sum_{i,j=1}^{n}|\Phi_{ij}|,$$
$$A(\Pi,\Phi)=\Pi-\Phi,\quad b=0,$$

and each problem instance considers a different value of $s$ which is part of the process of generating $\Sigma$.

We now present the numerical results for this set of problem instances in Tables 4.4 and 4.5.

| $s$ | Iteration Count | | | | | | |
|---|---|---|---|---|---|---|---|
| | ADL0 | ADL1 | ADL2 | iALM | IPL | QP | RQP |
| 5 | 2499 | 5131 | 5117 | 44050 | **1212** | 30783 | 2760 |
| 10 | 2486 | 6464 | 5475 | 45144 | **1184** | 32127 | 2827 |
| 15 | 2466 | 5526 | 5399 | 45985 | **1204** | 30784 | 2798 |

**Table 4.4:** (Innermost) Iteration Counts for SPCA problems.

| $s$ | Runtime (seconds) | | | | | | |
|---|---|---|---|---|---|---|---|
| | ADL0 | ADL1 | ADL2 | iALM | IPL | QP | RQP |
| 5 | 25.27 | 53.65 | 48.89 | 251.92 | **11.20** | 293.39 | 25.88 |
| 10 | 24.52 | 62.92 | 56.68 | 257.34 | **11.07** | 302.74 | 27.24 |
| 15 | 23.93 | 54.01 | 56.34 | 260.22 | **10.60** | 293.75 | 27.38 |

**Table 4.5:** Runtime for SPCA problems.

## 4.3 Linearly-Constrained Quadratic Matrix Problem

Given a pair of dimensions $(l,n)\in\mathbb{N}^2$, scalar pair $(\alpha_1,\alpha_2)\in\mathbb{R}^2_{++}$, linear operators $\mathcal{A}:S^n_+\mapsto\mathbb{R}^l$ , $\mathcal{B}:S^n_+\mapsto\mathbb{R}^n$, and $\mathcal{C}:S^n_+\mapsto\mathbb{R}^l$ defined by

$$[\mathcal{A}(z)]_i=\langle A_i,z\rangle\,,\quad [\mathcal{B}(z)]_j=\langle B_j,z\rangle\,,\quad [\mathcal{C}(z)]_i=\langle C_i,z\rangle\,,$$

for matrices $\{A_i\}_{i=1}^l,\{B_j\}_{j=1}^n,\{C_i\}_{i=1}^l\subseteq\mathbb{R}^{n\times n}$, positive diagonal matrix $D\in\mathbb{R}^{n\times n}$, and vector pair $(b,d)\in\mathbb{R}^l\times\mathbb{R}^l$, this subsection considers the following linearly-constrained quadratic matrix (LCQM) problem:

$$\min_z\ \frac{\alpha_1}{2}\|\mathcal{C}(z)-d\|^2-\frac{\alpha_2}{2}\|D\mathcal{B}(z)\|^2$$
$$\text{s.t. } \mathcal{A}(z)=b,\quad z\in P_n,$$

where $P_n=\{z\in S^n_+:\operatorname{tr} z=1\}$ denotes the $n$-dimensional spectraplex.

We now describe the experiment parameters for the instances considered. First, the dimensions are set to $(l,n)=(20,100)$ and only 1.0% of the entries of the submatrices $A_i,B_j$, and $C_i$ are nonzero. Second, the entries of $A_i,B_j,C_i,b$, and $d$ (resp., $D$) are generated by sampling from the uniform distribution $\mathcal{U}[0,1]$ (resp., $\mathcal{U}[1,1000]$). Third, the initial starting point $z_0$ is a random point in $S^n_+$. More specifically, three unit vectors $\nu_1,\nu_2,\nu_3\in\mathbb{R}^n$ and three scalars $e_1,e_2,e_2\in\mathbb{R}_+$ are first generated by sampling vectors $\tilde{\nu}_i\sim\mathcal{U}^n[0,1]$ and scalars $\tilde{d}_i\sim\mathcal{U}[0,1]$ and setting $\nu_i=\tilde{\nu}_i/\|\tilde{\nu}_i\|$

and $e_i = \tilde{e}_i/(\sum_{j=1}^3 \tilde{e}_i)$ for $i = 1, 2, 3$. The initial iterate for the first subproblem is then set to $z_0 = \sum_{i=1}^3 e_i \nu_i \nu_i^T$. Fourth, using the well-known fact that $\|z\|_F \leq 1$ for every $z \in P_n$, the auxiliary parameters for the iALM are

$$B_i = \|A_i\|_F, \quad L_i = 0, \quad \rho_i = 0 \quad \forall i \geq 1.$$

Finally, the composite form of the problem is

$$f(z) = \frac{\alpha_1}{2}\|\mathcal{C}(z) - d\|^2 - \frac{\alpha_2}{2}\|D\mathcal{B}(z)\|^2, \quad h(z) = \delta_{P_n}(z), \quad A(z) = \mathcal{A}(z),$$

and each problem instance uses a scalar pair $(\alpha_1, \alpha_2) \in \mathbb{R}^2_{++}$ so that $m = -\lambda_{\min}(\nabla^2 f) = 10$ and $M = \lambda_{\max}(\nabla^2 f)$ is a particular value given in the table below.

We now present the numerical results for this set of problem instances in Tables 4.6 and 4.7.

| $M$ | Iteration Count | | | | | | |
|---|---|---|---|---|---|---|---|
| | ADL0 | ADL1 | ADL2 | iALM | IPL | QP | RQP |
| $10^2$ | 638 | 4524 | 4339 | 43411 | **435** | 11178 | 829 |
| $10^3$ | 253 | 948 | 896 | 21479 | **208** | 5856 | 775 |
| $10^4$ | **386** | 2324 | 1632 | 30426 | 455 | 3659 | 663 |
| $10^5$ | 2445 | 7743 | 5113 | 78575 | 1511 | 11438 | **825** |
| $10^6$ | 8956 | 38904 | 23536 | 206881 | 6515 | 23833 | **947** |

**Table 4.6:** (Innermost) Iteration Counts for LCQM problems.

| $M$ | Runtime (seconds) | | | | | | |
|---|---|---|---|---|---|---|---|
| | ADL0 | ADL1 | ADL2 | iALM | IPL | QP | RQP |
| $10^2$ | 13.80 | 85.24 | 84.21 | 470.80 | **9.10** | 215.78 | 16.43 |
| $10^3$ | **4.39** | 18.40 | 17.31 | 232.41 | 4.53 | 117.03 | 15.81 |
| $10^4$ | **7.66** | 46.27 | 31.30 | 330.87 | 9.82 | 74.26 | 13.55 |
| $10^5$ | 49.38 | 153.89 | 102.10 | 824.18 | 29.28 | 212.44 | **16.03** |
| $10^6$ | 165.01 | 707.17 | 427.13 | 2091.60 | 119.97 | 422.90 | **16.25** |

**Table 4.7:** Runtimes for LCQM problems.

# 5 Concluding Remarks

This section gives some concluding remarks. For conciseness, we make our remarks under the assumption that $\hat{\rho} = \hat{\eta} = \varepsilon$ for some global tolerance $\varepsilon > 0$.

Similar to the analyses in [15, 18], the analysis of the AIDAL method strongly makes use of assumption (A3) and the assumption that $D_h < \infty$ to obtain its competitive $\mathcal{O}(\varepsilon^{-5/2} \log \varepsilon^{-1})$ iteration complexity. However, we conjecture that the these two assumptions may be removed using the more complicated analysis in [22] to obtain a slightly worse $\mathcal{O}(\varepsilon^{-3} \log \varepsilon^{-1})$ iteration complexity (like in [22]).

We now give some remarks about the numerical experiments. Before proceeding, we first summarize the basic details of the tested methods — which all can be described as inexact augmented Lagrangian methods under specific dampening and under-relaxation parameters $\theta$ and $\chi$ — in Table 5.1. Besides these details, we also state some more nuanced properties about the methods. First,

| Name | Warm-start? | Line Search? | | Parameter Values | | Complexity |
|---|---|---|---|---|---|---|
| | $z_k$ | $\lambda$ | $M$ | $\theta$ | $\chi$ | $\varepsilon = \hat{\rho} = \hat{\eta}$ |
| ADL$k$ | ✓ | ✘ | ✓ | $[0, 1)$ | $(0, 1]$ | $\mathcal{O}(\varepsilon^{-5/2} \log \varepsilon^{-1})$ |
| iALM | ✓ | ✘ | ✘ | 0 | $(0, 1]$ | $\mathcal{O}(\varepsilon^{-5/2} \log \varepsilon^{-1})$ |
| IPL | ✓ | ✘ | ✓ | 0 | 1 | $\mathcal{O}(\varepsilon^{-3} \log \varepsilon^{-1})$ |
| QP | ✘ | ✘ | ✓ | 1 | 0 | $\mathcal{O}(\varepsilon^{-3})$ |
| RQP | ✓ | ✓ | ✓ | 1 | 0 | $\mathcal{O}(\varepsilon^{-3} \log \varepsilon^{-1})$ |

**Table 5.1:** *Comparison of tested methods.* The first column lists the names of the methods at the beginning of the section. The second one indicates whether or not a method uses the last obtained approximate solution of a prox subproblem as the starting point in the solver of the next prox subproblem. The third (resp. fourth) one indicates whether a line search is used to obtain the prox stepsize parameter $\lambda$ (resp. estimate the the curvature constant $M$ in assumption (A2)) for each of its prox subproblems. The fifth and sixth columns indicate the range of values for the under-relaxation parameters $\chi$ and $\theta$. The last column gives the (innermost) iteration complexity of the methods.

the iALM differs from the other tested methods in that it uses an ACG variant with a termination criteria that is different from the one in (46) and/or its relaxation. Second the RQP method differs from the ADL$k$, QP, and IPL methods in that it uses an ACG variant with a more relaxed version of (46) which allows for nonconvex prox subproblems. Third, the main difference between the ADL$k$ and IPL methods is in how they decide when to double $c_k$, i.e., step 3 of Algorithm 2.1. In particular, the condition used in the IPL method depends on both $\nu$ and $k$ whereas the condition in the ADL$k$ methods does not. Finally, it is worth mentioning that the QP method is the only method that can be run without requiring any regularity conditions on the linear constraint and without assuming that $D_h < \infty$.

Keeping the above properties in mind, we now give the promised remarks about our numerical experiments. First, across different values of $M$, the most stable method appears to be the RQP method due to its use of line search subroutines to estimate $\lambda$ and $M$. Second, the IPL and ADL0 methods perform similarly as they only differ in how they update the penalty parameter $c_k$. Finally, other than the RQP method, all methods appear to suffer when the curvature ratio $M/m$ is too small.

In view of these promising results, a future avenue of research is the development of a proximal augmented Lagrangian method in which $\lambda$ is chosen potentially larger than $1/m$ and in an adaptive manner (like in the R-QP-AIPP of [10]).

# A Statement and Properties of the ACG Algorithm

Recall from Section 1 that our interest is in solving (1) by inexactly solving NCO subproblems of the form in (3). This subsection presents an ACG algorithm for inexactly solving latter type of problem and it considers the more general class of NCO problems

$$\min_{u \in \mathbb{R}^n} \{\psi_s(u) + \psi_n(u)\}, \tag{45}$$

where the functions $\psi_s$ and $\psi_n$ are assumed to satisfy the following assumptions:

(B1) $\psi_n : \mathbb{R}^n \mapsto (-\infty, \infty]$ is a proper, closed, and $\mu$-strongly convex function with $\mu > 0$.

(B2) $\psi_s$ is continuously differentiable on $\mathbb{R}^n$ and satisfies $\psi_s(z') - \ell_{\psi_s}(z'; z) \leq L\|z' - z\|^2/2$ for some $L \geq 0$ and every $z', z \in \mathbb{R}^n$.

Clearly, problem (3) is a special case of (45), and hence, any result that is stated in the context of (45) also applies to (3).

We first state the ACG algorithm in Algorithm A.1 which, for a given a pair $(\tilde{\sigma}, x_0) \in \mathbb{R}_{++} \times \operatorname{dom}\psi_n$, inexactly solves (45) by obtaining a triple $(z, v, \varepsilon)$ satisfying

$$v \in \partial_\varepsilon(\psi_s + \psi_n)(z), \quad \|v\|^2 + 2\varepsilon \le \tilde{\sigma}^2 \|x_0 - z + v\|^2. \tag{46}$$

Note that if ACG algorithm obtains the aforementioned triple with $\sigma = 0$ then the first component of the triple is, in fact, a global solution of (45).

**Algorithm A.1:** Accelerated Composite Gradient (ACG) Algorithm

**Require:** $(\tilde{\sigma}, x_0) \in (0,1) \times \operatorname{dom}\psi_n$.
**Output :** a triple $(z, v, \varepsilon) \in \operatorname{dom}\psi_n \times \mathbb{R}^n \times \mathbb{R}_{++}$ satisfying (46).

**Function** `AIDAL`$([\psi_s, \psi_n], [L, \mu], \tilde{\sigma}, x_0)$**:**
- `STEP 0` (initialization):
- Set $y_0 \leftarrow x_0$, $A_0 \leftarrow 0$, $\Gamma_0 \leftarrow 0$.
- **for** $j \leftarrow 1, 2, \ldots$ **do**
  - `STEP 1` (main iterates):
  - $A_{j+1} \leftarrow A_j + \frac{\mu A_j + 1 + \sqrt{(\mu A_j + 1)^2 + 4L(\mu A_j + 1)A_j}}{2L}$
  - $\tilde{x}_j \leftarrow \frac{A_j}{A_{j+1}} x_j + \frac{A_{j+1} - A_j}{A_{j+1}} y_j$
  - $\Gamma_{j+1}(\cdot) \leftarrow \frac{A_j}{A_{j+1}} \Gamma_j(\cdot) + \frac{A_{j+1} - A_j}{A_{j+1}} \ell_{\psi_s}(\cdot, \tilde{x}_j)$ ▷ Note that $\Gamma_{j+1}$ is affine
  - $y_{j+1} \leftarrow \operatorname{argmin}_{y \in \mathbb{R}^n} \left\{ \Gamma_{j+1}(y) + \psi_n(y) + \frac{1}{2A_{j+1}} \|y - y_0\|^2 \right\}$
  - $x_{j+1} \leftarrow \frac{A_j}{A_{j+1}} x_j + \frac{A_{j+1} - A_j}{A_{j+1}} y_{j+1}$
  - `STEP 2` (auxiliary residuals):
  - $u_{j+1} \leftarrow \frac{y_0 - y_{j+1}}{A_{j+1}}$
  - $\eta_{j+1} \leftarrow \psi(x_{j+1}) - \Gamma_{j+1}(y_{j+1}) - \psi_n(y_{j+1}) - \langle u_{j+1}, x_{j+1} - y_{j+1} \rangle$;
  - `STEP 3` (termination check):
  - **if** $\|u_j\|^2 + 2\eta_j \le \tilde{\sigma}^2 \|u_j + y_0 - y_j\|^2$ **then**
    - **return** $(y_j, u_j, \eta_j)$ ▷ Stop and output

The below result, whose proof can be found in [11, Propostion 2.1], states some basic properties about the algorithm.

**Proposition A.1.** *The following properties hold about the ACG algorithm:*

*(a) for every $j \ge 1$, we have*

$$u_j \in \partial_{\eta_j}(\psi_s + \psi_n)(x_j), \quad A_j \ge \frac{1}{L} \max\left\{ \frac{j^2}{4}, \left(1 + \sqrt{\frac{\mu}{4L}}\right)^{2(j-1)} \right\};$$

*(b) for any $\tilde{\sigma} > 0$, the ACG algorithm stops and outputs a triple $(z, v, \varepsilon)$ satisfying*

$$v \in \partial_\varepsilon(\psi_s + \psi_n)(z), \quad \|v\|^2 + 2\varepsilon \le \tilde{\sigma}^2 \|v + x_0 - z\|^2 \tag{47}$$

*in at most*

$$\left\lceil 1 + \sqrt{\frac{L}{\mu}} \log_1^+ \left( \left\lceil \frac{1 + \tilde{\sigma}}{\tilde{\sigma}} \right\rceil \sqrt{2L} \right) \right\rceil$$

*ACG iterations.*

## B Technical Inequalities

This appendix gives some technical inequalities that will be necessary in the main body of the paper.

The first result, whose proof can be found in [4, Lemma 1.3], presents a relationship between elements in the image of a linear operator.

**Lemma B.1.** *For any $S \in \mathbb{R}^{m\times n}$ and $u \in \operatorname{Im} S$, we have $\sigma_S^+\|u\| \le \|Su\|$.*

The next result is essential in proving Proposition 3.3 in Section 3.

**Lemma B.2.** *For any $(\tau,\theta) \in [0,1]^2$ satisfying $\tau \le \theta^2$ and any $a, b \in \mathbb{R}^n$, we have that*

$$\|a-(1-\theta)b\|^2 - \tau\|a\|^2 \ge \left[\frac{(1-\tau)-(1-\theta)^2}{2}\right]\left(\|a\|^2-\|b\|^2\right). \tag{48}$$

*Proof.* Let $a, b \in \mathbb{R}^n$ be fixed and define

$$z = \begin{bmatrix} \|a\| \\ \|b\| \end{bmatrix}, \quad M = \begin{bmatrix} (1-\tau)+(1-\theta)^2 & -2(1-\theta) \\ -2(1-\theta) & (1-\tau)+(1-\theta)^2 \end{bmatrix}. \tag{49}$$

Moreover, using our assumption of $\tau \le \theta^2 \le 1$, observe that

$$\begin{aligned} \det M &= \left[(1-\tau)+(1-\theta)^2-2(1-\theta)\right]\left[(1-\tau)+(1-\theta)^2+2(1-\theta)\right] \\ &= \left[\theta^2-\tau\right]\left[(1-\tau)+(1-\theta)^2+2(1-\theta)\right] \ge 0, \end{aligned}$$

and hence, by Sylvester's criterion, it follows that $M \succeq 0$. Combining this fact with the Cauchy-Schwarz inequality and (49), we thus have that

$$\begin{aligned} &\|a-(1-\theta)b\|^2-\tau\|a\|^2 \ge (1-\tau)\|a\|^2 - 2(1-\theta)\|a\|\cdot\|b\| + (1-\theta)^2\|b\|^2 \\ &= \frac{1}{2}z^TMz + \left[\frac{(1-\tau)-(1-\theta)^2}{2}\right]\left(\|a\|^2-\|b\|^2\right) \ge \left[\frac{(1-\tau)-(1-\theta)^2}{2}\right]\left(\|a\|^2-\|b\|^2\right). \end{aligned}$$

□

## C Approximate Subdifferential

This appendix contains results related to the approximate subdifferential.

The proof of the below result can be found, for example, in [21, Lemma A.2].

**Lemma C.1.** *Let a proper function $\psi : \mathbb{R}^n \mapsto (-\infty, \infty]$, scalar $\sigma \in (0,1)$, and $(z, z^-) \in \operatorname{dom}\psi \times \mathbb{R}^n$ be given, and assume there exists $(v, \varepsilon)$ such that*

$$v \in \partial_\varepsilon\left(\psi + \frac{1}{2}\|\cdot - z^-\|^2\right)(z), \quad \|v\|^2 + 2\varepsilon \le \sigma^2\|v+z^- - z\|^2.$$

*Then, for every $u \in \mathbb{R}^n$ and $s > 0$, we have*

$$\psi(z) + \frac{1}{2}\left[1-\sigma^2\left(\frac{1+s}{s}\right)\right]\|v+z^- - z\|^2 \le \psi(u) + \left(\frac{1+s}{2}\right)\|u-z^-\|^2.$$

The proof of the next result can be found, for example, in [17, Lemma 3.4].

**Lemma C.2.** *Suppose $\psi \in \overline{\text{Conv}}\ \mathbb{R}^n$ is a $\xi$-strongly convex function with respect to a norm $\|\cdot\|_{\mathcal{Z}}$ and let $(y,\eta) \in \mathbb{R}^n \times \mathbb{R}$ be such that $0 \in \partial_\eta \psi(y)$. Then,*

$$0 \in \partial_{2\eta}\left(\psi - \frac{\xi}{4}\|\cdot - y\|_{\mathcal{Z}}^2\right)(y).$$

Finally, the proof of the last two results can be found in [21].

**Lemma C.3.** *Suppose $\psi \in \overline{\text{Conv}}\ \mathbb{R}^n$ is $K_\psi$-Lipschitz continuous. Then, for every $y \in \operatorname{dom}\psi$ and $\varepsilon \geq 0$, we have*

$$\partial_\varepsilon \psi(z) \subseteq \mathcal{B}(0; K_\psi) + N^\varepsilon_{\operatorname{dom}\psi}(z).$$

**Lemma C.4.** *Suppose $\psi \in \overline{\text{Conv}}\ \mathbb{R}^n$ is $K_\psi$-Lipschitz continuous with finite diameter $D_\psi$. Then, for every $\varepsilon \geq 0$, $y, \bar{y} \in \operatorname{dom} h$ and $\xi \in \partial_\varepsilon \psi(y)$, we have*

$$\|\xi\| \text{dist}_{\partial(\operatorname{dom}\psi)}(\bar{y}) \leq \left[\text{dist}_{\partial(\operatorname{dom}\psi)}(\bar{y}) + \|y - \bar{y}\|\right] K_\psi + \langle \xi, y - \bar{y}\rangle + \varepsilon.$$

The final result presents some variational properties about a single composite/proximal gradient step.

**Lemma C.5.** *Let $\psi_n \in \overline{\text{Conv}}\ \mathbb{R}^n$, $z \in \operatorname{dom}\psi_n$, and $\psi_s$ be a differentiable function on $\operatorname{dom} h$ satisfying*

$$\|\nabla \psi_s(x) - \nabla \psi_s(y)\| \leq L\|x - y\| \quad \forall x, y \in \operatorname{dom}\psi_n, \tag{50}$$

*for some $L \geq 0$. Moreover, define the quantities*

$$\hat{z} := \underset{u}{\operatorname{argmin}} \left\{ \ell_{\psi_s}(u; z) + \psi_n(u) + \frac{L}{2}\|u - z\|^2 \right\}, \tag{51}$$

$$w := L(z - \hat{z}), \quad \hat{v} := w + \nabla\psi_s(\hat{z}) - \nabla\psi_s(z), \tag{52}$$

*If $0 \in \partial_\varepsilon(\psi_s + \psi_n)(z)$, for some $\varepsilon \geq 0$, then*

$$w \in \nabla\psi_s(z) + \partial_\varepsilon\psi_n(z), \quad \hat{v} \in \nabla\psi_s(\hat{z}) + \partial\psi_n(\hat{z}), \tag{53}$$

$$\|z - \hat{z}\| \leq \sqrt{\frac{2\varepsilon}{L}}, \quad \|w\| \leq \sqrt{2L\varepsilon}, \quad \|\hat{v}\| \leq 2\sqrt{2L\varepsilon}. \tag{54}$$

*Proof.* It is well-known that (50) implies $\psi_s(x) - \ell_{\psi_s}(x; y) \leq L\|x - y\|^2/2$ for every $x, y \in \operatorname{dom}\psi_n$. In view of the previous set of inequalities and the fact that $0 \in \partial_\varepsilon(\psi_s + \psi_n)(z)$ implies

$$(\psi_s + \psi_n)(z) - (\psi_s + \psi_n)(\hat{z}) \leq \varepsilon, \tag{55}$$

it follows from [9, Lemma 19] that (53) holds. To show the desired inclusions, we first use (55) with [9, Lemma 19] to obtain

$$\|w\|^2 = L^2\|z - \hat{z}\|^2 \leq 2L\left[(\psi_s + \psi_n)(z) - (\psi_s + \psi_n)(\hat{z})\right] \leq 2L\varepsilon, \tag{56}$$

which clearly implies the first two inequalities in (55). It now follows from (56), the definition of $\hat{v}$, (50), and the triangle inequality that

$$\|\hat{v}\| \leq \|w\| + \|\nabla\psi_s(\hat{z}) - \nabla\psi_s(z)\| \leq \sqrt{2L\varepsilon} + L\|z - \hat{z}\| \leq 2\sqrt{2L\varepsilon}$$

which is exactly the second inequality in (55). □

## References


[1] N. S. Aybat and G. Iyengar. A first-order smoothed penalty method for compressed sensing. *SIAM J. Optim.*, 21(1):287–313, 2011.

[2] N. S. Aybat and G. Iyengar. A first-order augmented Lagrangian method for compressed sensing. *SIAM J. Optim.*, 22(2):429–459, 2012.

[3] D. Boob, Q. Deng, and G. Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Available on arXiv:1908.02734*, 2019.

[4] M. L. N. Goncalves, J. G. Melo, and R. D. C. Monteiro. Convergence rate bounds for a proximal admm with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems. *Avaliable on arXiv:1702.01850*, 2017.

[5] Q. Gu, Z. Wang, and H. Liu. Sparse pca with oracle property. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Adv. Neural Inf. Process. Syst. 27*, pages 1529–1537. Curran Associates, Inc., 2014.

[6] D. Hajinezhad and M. Hong. Perturbed proximal primal-dual algorithm for nonconvex nonsmooth optimization. *Math. Program.*, 176:207–245, 2019.

[7] B. Jiang, T. Lin, S. Ma, and S. Zhang. Structured nonconvex and nonsmooth optimization algorithms and iteration complexity analysis. *Comput. Optim. Appl.*, 72(3):115–157, 2019.

[8] W. Kong. Accelerated Inexact First-Order Methods for Solving Nonconvex Composite Optimization Problems. *Available on arXiv:2104.09685*, April 2021.

[9] W. Kong, J. G. Melo, and R. D. C. Monteiro. Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs. *SIAM J. Optim.*, 29(4):2566–2593, 2019.

[10] W. Kong, J. G. Melo, and R. D. C. Monteiro. An efficient adaptive accelerated inexact proximal point method for solving linearly constrained nonconvex composite problems. *Comput. Optim. Appl.*, 76(2):305–346, 2020.

[11] W. Kong, J. G. Melo, and R. D. C. Monteiro. Iteration-complexity of a proximal augmented lagrangian method for solving nonconvex composite optimization problems with nonlinear convex constraints. *Available on arXiv:2008.07080*, 2020.

[12] W. Kong and R. D. C. Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *Available on arXiv:1905.13433v2*, 2019.

[13] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order penalty methods for convex programming. *Math. Program.*, 138(1):115–139, Apr 2013.

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

[15] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu. Rate-improved inexact augmented Lagrangian method for constrained nonconvex optimization. *Available on arXiv:2007.01284*, 2020.

[16] Z. Li and Y. Xu. First-order inexact augmented Lagrangian methods for convex and nonconvex programs: nonergodic convergence and iteration complexity. *Personal technical report*, 2019.

[17] J. Liang and R. D. C. Monteiro. A Doubly Accelerated Inexact Proximal Point Method for Nonconvex Composite Optimization Problems. *Available on arXiv:1811.11378*, November 2018.

[18] Q. Lin, R. Ma, and Y. Xu. Inexact proximal-point penalty methods for constrained non-convex optimization. *Available on arXiv:1908.11518*, 2019.

[19] Y. F. Liu, X. Liu, and S. Ma. On the non-ergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming. *Available on arXiv:1603.05738*, 2016.

[20] Z. Lu and Z. Zhou. Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. *Available on arXiv:1803.09941*, 2018.

[21] J. G. Melo and R. D. C Monteiro. Iteration-complexity of an inner accelerated inexact proximal augmented Lagrangian method based on the classical lagrangian function and a full Lagrange multiplier update. *Available on arXiv:2008.00562*, 2020.

[22] J. G. Melo, R. D. C. Monteiro, and H. Wang. Iteration-complexity of an inexact proximal accelerated augmented Lagrangian method for solving linearly constrained smooth nonconvex composite optimization problems. *Available on arXiv:2006.08048*, 2020.

[23] R. D. C. Monteiro, C. Ortiz, and B. F. Svaiter. An adaptive accelerated first-order method for convex optimization. *Comput. Optim. Appl.*, 64:31–73, 2016.

[24] I. Necoara, A. Patrascu, and F. Glineur. Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming. *Optim. Methods Softw.*, pages 1–31, 2017.

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

[26] M. Sahin, A. Eftekhari, A. Alacaoglu, F. Latorre, and V. Cevher. An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints. *Available on arXiv:1906.11357*, 2019.

[27] Y. Xu. Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Available on arXiv:1711.05812*, 2017.

[28] J. Zhang and Z.-Q. Luo. A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. *Available on arXiv:2006.16440*, 2020.

[29] J. Zhang and Z.-Q. Luo. A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM J. Optim.*, 30(3):2272–2302, 2020.